COPERNICUS MASTER IN DIGITAL EARTH

PARIS LODRON UNIVERSITÄT SALZBURG

SPATIAL STATISTICS

SUMMARY OF ASSIGNMENTS

Instructor: Christian Neuwirth

Contents

Descriptive Statistics

Assignment 1

1. *Open Excel and generate a random, normally distributed sample of size n = 1000 (Mean = 50; Standard Deviation = 10) using the following equation:*

=NORM.INV(RAND();50;10)

Calculate the range, 1st quartile, 3rd quartile and interquartile range of this sample (Excel function for quartile =QUARTILE.EXKL()).

Range	1st Quartile	3rd Quartile	Interquartile Range
69.61	43.86	57.35	13.49

2. *Find the grouped mean of the following data:*

Category	Category Mean	Frequency
<20	10	5
20-39.99	30	15
40-59.99	50	10
60-79.99	70	12

Grouped mean: 43.81

3. *Ten migration distances corresponding to the distances moved by recent migrants are observed (in km): 43, 6, 7, 11, 122, 41, 21, 17, 1, 3. Find the mean and standard deviation. Interpret the result in your own words.*

Mean: 27.2

Standard deviation: 36.46

The mean of 27.2 km here indicates the average migration distance per migrant. At the same time, the standard deviation provides information about the high variability of the distance per migrant, as the average deviation of 36.46 km per migrant is quite high. This high value is partly due to the migrant with a migration distance of 122 km, since it differs greatly from the other migrants.

4. *Describe the following data by using common data categories (nominal, ordinal, dichotomous, discrete, etc.):*

- Intelligence scores → Interval Scale, continuous data
- Temperature in Kelvin or Temperature in °C → in Kelvin Proportional Scale; in °C Interval Scale; continuous data in both cases
- Rainy days in September → Proportional scale, discrete data
- Hair color (blond, black, brown) → Nominal Scale, discrete data
- Rankings of teams in a tournament → Rank Scale, discrete data
- Body size classified as tiny, small, medium, tall, and giant → Ordinal Scale, discrete data
- Customer satisfaction (satisfied or not satisfied) → Nominal Scale, dichotomous data

5.

a) *Create a data frame containing coordinates and population numbers of nine Austrian federal capitals.*

City	**X**	**Y**	**Population**
Wien	16.3738	48.2082	2005760
Graz	15.4395	47.0707	302749
Linz	14.2858	48.3064	211944
Salzburg	13.0550	47.8095	157399
Innsbruck	11.3923	47.2682	132188
Klagenfurt	14.3122	46.6365	104866
St. Pölten	15.6333	48.2000	58856
Bregenz	9.7471	47.5031	29643
Eisenstadt	16.5246	47.8457	16037

b) *Find the unweighted mean center of cities*

(X/Y) Coordinates → **(14.08/47.65)**

c) *Find the population weighted mean center of cities.*

(X/Y) Coordinates → **(15.59/47.98)**

d) Add centers to a leaflet basemap

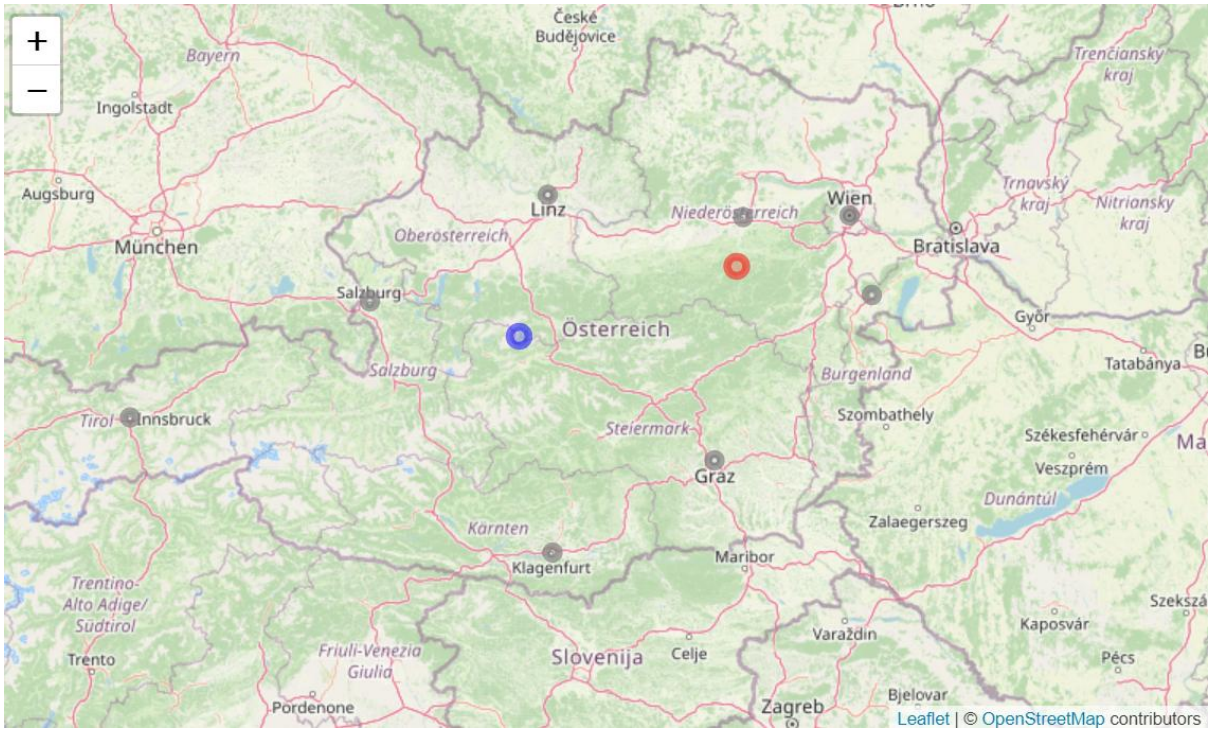

Figure 1. Map showing the Austrian federal capitals (grey), their unweighted mean center (blue) & the weighted mean center by population (red).

Exploratory (Spatial) Data

Assignment 2

1. *Get the Top 15 Ukrainian agricultural production items in 2023 with respect to production in tons*
   - *Select relevant columns (dplyr::select)*
   - *Filter relevant rows by means of a condition (dplyr::filter)*
   - *Sort production items from highest to lowest production value (dplyr::arrange)*
   - *Slice table and only include the top 15 Ukrainian production items (dplyr::slice_head)*
   - *and render results in a table (knitr::kable)*

Table: Top 15 Ukrainian Agricultural Products in 2023 by Production (tons)

Area	Item	Element	Y2023
Ukraine	Cereals, primary	Production	59307682
Ukraine	Maize (corn)	Production	31030440

```
|Ukraine |Wheat                     |Production | 21625170|
|Ukraine |Potatoes                  |Production | 21358630|
|Ukraine |Roots and Tubers, Total   |Production | 21358630|
|Ukraine |Sugar beet                |Production | 13129710|
|Ukraine |Sugar Crops Primary       |Production | 13129710|
|Ukraine |Sunflower seed            |Production | 12759690|
|Ukraine |Oilcrops, Cake Equivalent |Production | 12345563|
|UkraiSne |Hen eggs in shell, fresh  |Production | 11265600|
|Ukraine |Vegetables Primary        |Production |  8301542|
|Ukraine |Oilcrops, Oil Equivalent  |Production |  7727996|
|Ukraine |Milk, Total               |Production |  7430400|
|Ukraine |Raw milk of cattle        |Production |  7267100|
|Ukraine |Barley                    |Production |  5507190|
```

2. *Choose one of the Top 5 Ukrainian production items and retrieve the national productions in tons of this item in 2023 (Y2023) from the Europe-wide table (in a new RScript). Create a histogram to visualize the quantitative distribution of that variable. Find a suitable binwidth and interpret the result (e.g. data is left skewed – many countries produce large amounts of XY, few countries produce little…, the Ukraine produces the majority…)*

Disclaimer: I have chosen **potatoes** as the production item, which is the fourth most important product in Ukraine with 21358630 tons in 2023.

To effectively visualize the quantitative distribution of the potato production in Europe with a histogram, I have chosen a bandwidth of 500000. This allows for the interpretation of the result, which shows that the data is clearly right skewed, indicating that most European countries produce only small quantities of potatoes. At this point it should also be noted that many of these countries with low production rates could also be small in area size and should therefore better not be compared with larger countries. However, there are only a few outliers with very high potato production.

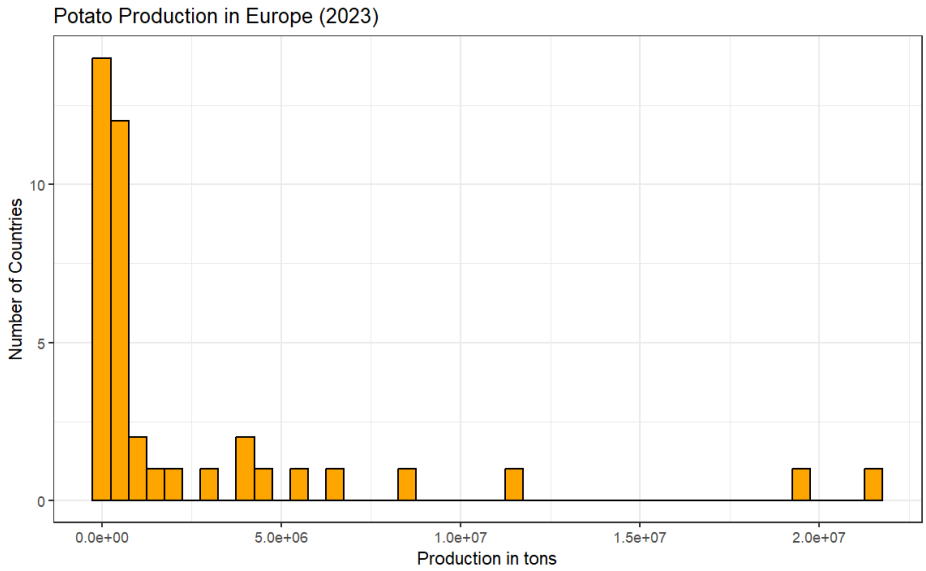

Figure 1. Histogram showing the potato production (tons) in Europe in 2023.

To better understand which countries these outliers with high production rates are, I have additionally created a bar plot showing the Top 15 European countries in potato production (tons) in 2023. There you can see that the highest production is in Ukraine, followed by the Russian Federation, which are probably the countries shown in the histogram at the right-hand side. In addition to these two countries, Germany and France, both European countries with a large area size, also have a high potato production.

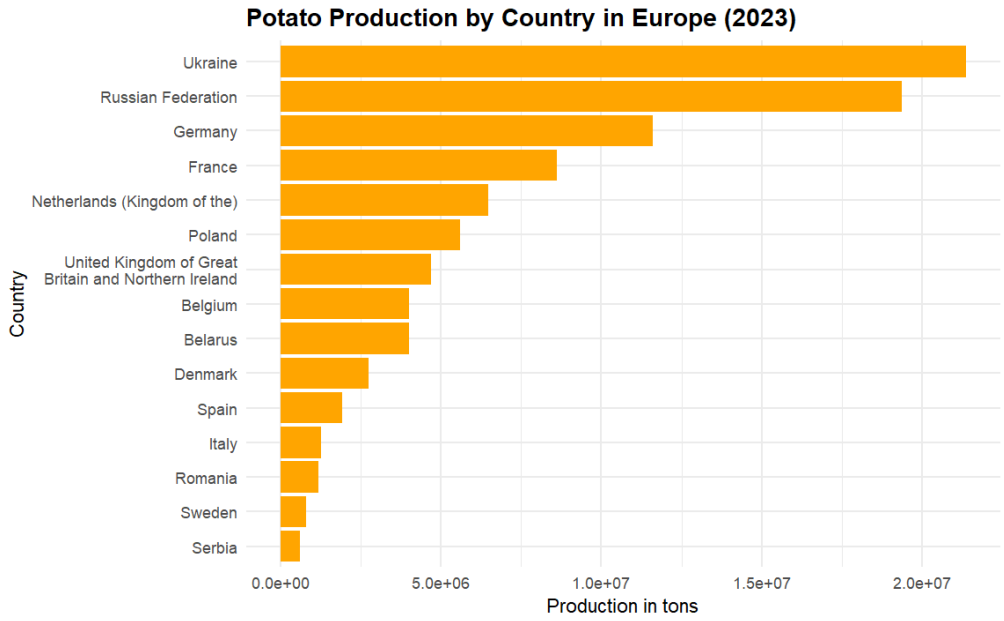

Figure 2. Barplot showing the Top 15 European countries regarding their potato production (tons) in 2023.

3. *Figure 2 in this Article shows changes in national mean body mass (BMI) by means of a QQ-Plot: For instance, the body mass index in Nigeria increased by 0.031 between 2003 and 2008. The QQ-Plot shows how the distribution of the data in Nigeria changed between 2003 and 2008*
   •Interpret one of the countries in Figure 2
   • How did the distribution in the data change?
   • What information is in the QQ-plot that is not apparent from changes in mean BMI alone?

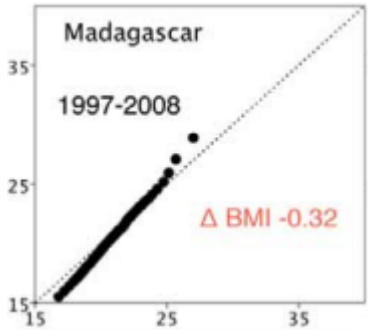

Figure 3. QQ plot of BMI in Madagascar (Source: Razak et al., 2013; doi:10.1371/journal.pmed.1001367)

The QQ plot shows the distribution of BMI in Madagascar between 1997 and 2008, with the x-axis representing the baseline survey (1997) and the y-axis the final survey (2008). The dots represent a quantile of the BMI in 1997 compared to the corresponding quantile in 2008. At the same time, the dotted diagonal line in the center of the diagram represents the equality between the baseline & the final survey and thus indicates the status of no change. In addition to the QQ plot, the change in the mean BMI between 1997 and 2008 is also shown. The red color here indicates a decrease in the mean BMI of -0.32, which means that the BMI has worsened on average in the defined period.

Since the mean BMI difference is only a single value that hides important variation in the data, the QQ plot can be used additionally to show how different parts of the population were affected. For instance, it can be seen that a significant proportion of the population that had a low BMI in 1997 has an even lower BMI in 2008. At the same time, the opposite trend can also be observed: People with a high BMI in 1997 gain even more weight in 2008. It should be noted here that the density of dots for the second trend is rather low, which might indicate outliers. In conclusion, the QQ plot confirms the mean of the BMI difference by showing that the population with a low BMI has worsened in terms of body weight, while it also shows the variation of the BMI development in Madagascar with few people increasing their BMI between 1997 and 2008.

Inferential Statistics

Assignment 3

1. *In Excel (alternatively you may use R!) use function =NORM.S.DIST(t, TRUE) to calculate empirical p-values. The output of the function corresponds to the area under the probability density function to the left of a threshold value. Calculate the p-value for a two-sided hypothesis.*

I did this exercise using R and the function pnorm() instead of NORM.S.DIST(t,TRUE)

```
#Average yearly precipitation
x1 <- 1100
x2 <- 1050
#Sample size
n1 <- 105
n2 <- 150
#Standard Deviation
s1 <- 125
s2 <- 100
#Pooled standard deviation
sp <- sqrt((s1^2 + s2^2) / 2)
#Standard error of the difference
se_diff <- sp * sqrt(1/n1 + 1/n2)
#Difference D
D<-x1-x2
#Significance test
t<-D/se_diff
#Degrees of freedom
df<-n1-1+n2-1
#Calculate empirical p-value
p_value<-2*(1-pnorm(abs(t)))
#Significance level
alpha <- 0.05
if (p_value < alpha) {
  cat("The result is significant (p <", alpha, ")\n")
} else {
  cat("The result is not significant (p >=", alpha, ")\n")
}
```

Result (p-value for a two-sided hypothesis): 0.000517

2. *Specify a significance level. Does the test produce significant results? Interpret the result in your own words.*

I specified a significance level of 0.05 to determine whether the test provides significant results. Since the p-value of 0.000517 is below the significance level of 0.05, the result is significant. This means that we reject H_0, which states that there is no difference in the average precipitation values between the two locations. Based on the t-test, we have evidence that the observed difference in precipitation is

unlikely to be due to random chance and is statistically significant, potentially indicating the influence of different geographic conditions.

Spatial Random Distributions

3. *Draw a rectangle that is 12cm by 10cm on a sheet of paper. Locate 30 dots at random within the rectangle. This means that every dot should be located independently of the other dots! Then draw a six-by-five grid of 30 squares cells on top of your rectangle. You can do this by making little tick marks at 2cm intervals along the side of your rectangle. Connecting the tick marks will divide your original rectangle into 30 squares, each having a side length of 2cm. Sum up cell values according to the following schema: cell without dots -> cell value 1, one dot in cell -> cell value 0, two dots in cell -> cell value 1, three dots in cell -> cell value 4, four dots in cell -> cell value 9, five dots in cell -> cell value 16, six dots in cell -> cell value 25, seven dots in cell -> cell value 36*

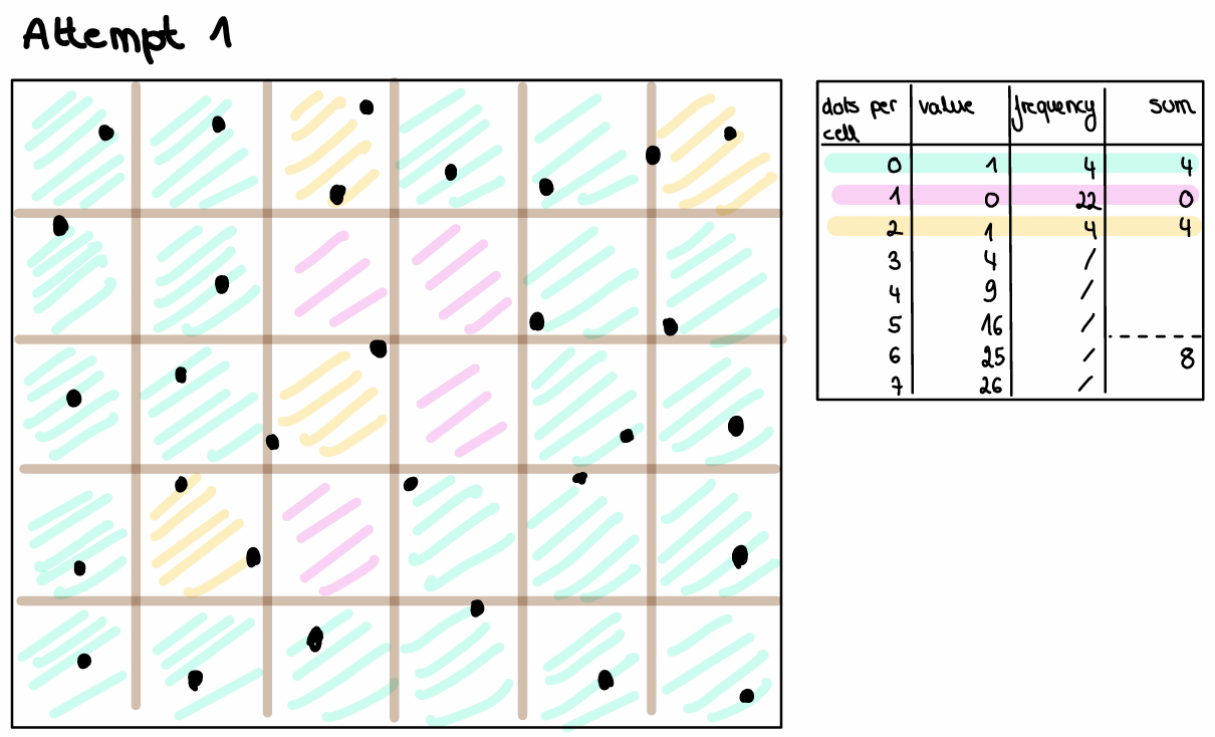

Figure 1. First attempt of creating a random distribution of 30 dots in a 6x5 grid.

Although the dots created in the first attempt were aimed at a random distribution, calculating the sum according to the previously defined schema showed that the values are not randomly distributed, but exhibit uniformity.

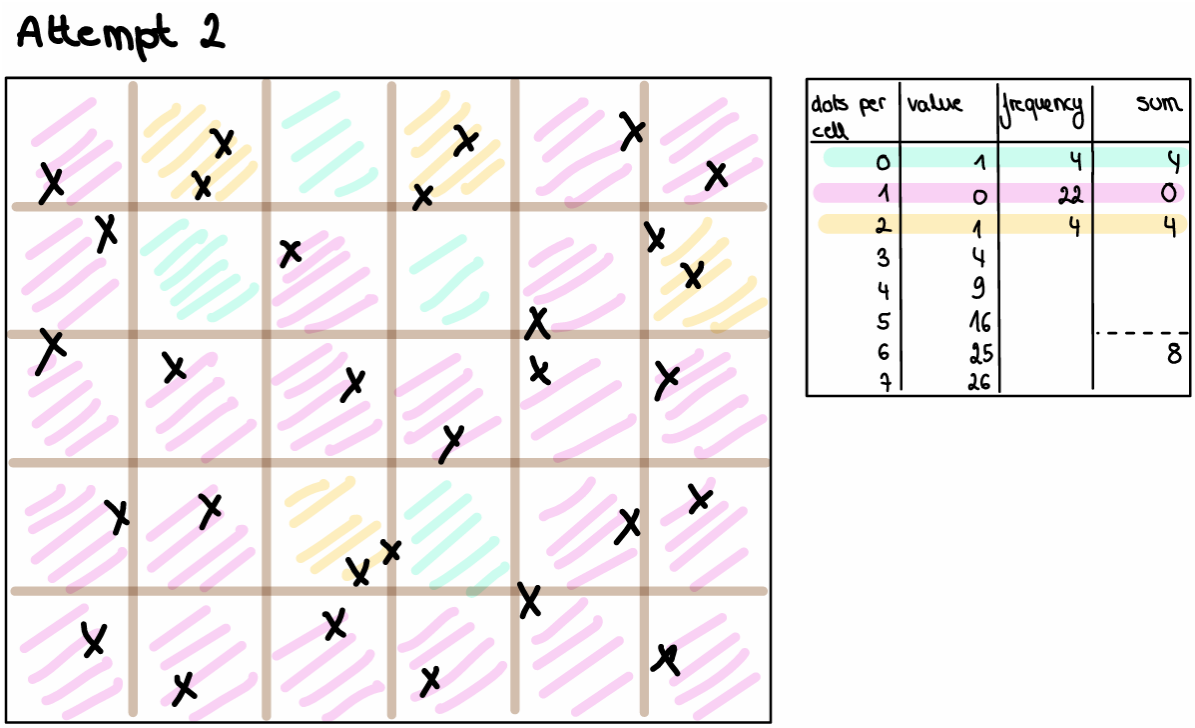

Figure 2. Second attempt of creating a random distribution of 30 dots in a 6x5 grid.

Since the first attempt did not produce a random distribution, I made a second attempt, which produced exactly the same sum as the first, although the distribution in the cells was slightly different.

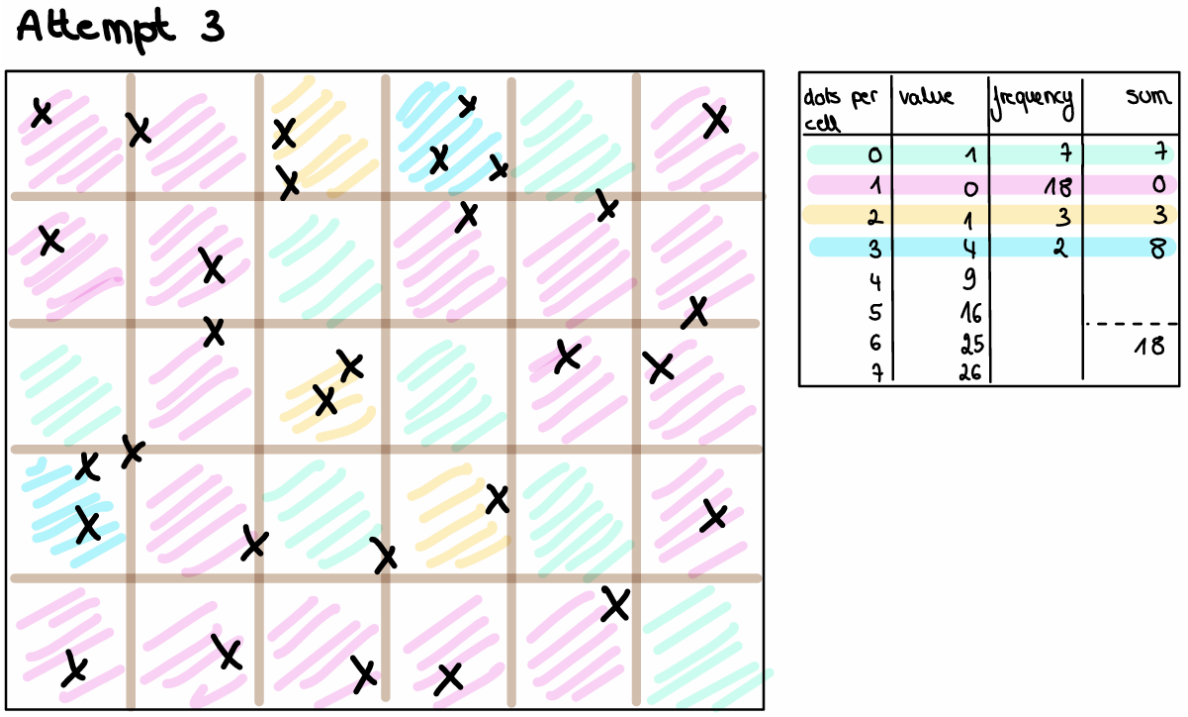

Figure 3. Third attempt of creating a random distribution of 30 dots in a 6x5 grid.

In the third attempt, I concentrated on keeping the points close together, but still avoiding spatial clustering. Using this technique, I was able to achieve a random distribution for the first time, although the sum of 18 is only just above the defined threshold of 17 for a normal distribution. In conclusion, these three attempts have shown how difficult it can be to create a random distribution by hand, which is why this task is often performed by a machine.

4. *Create a random point distribution in R. Run the R Script multiple times and answer the following questions. Is the result probabilistic or deterministic? Is there only one or many realizations of a spatial random distribution / H0?*

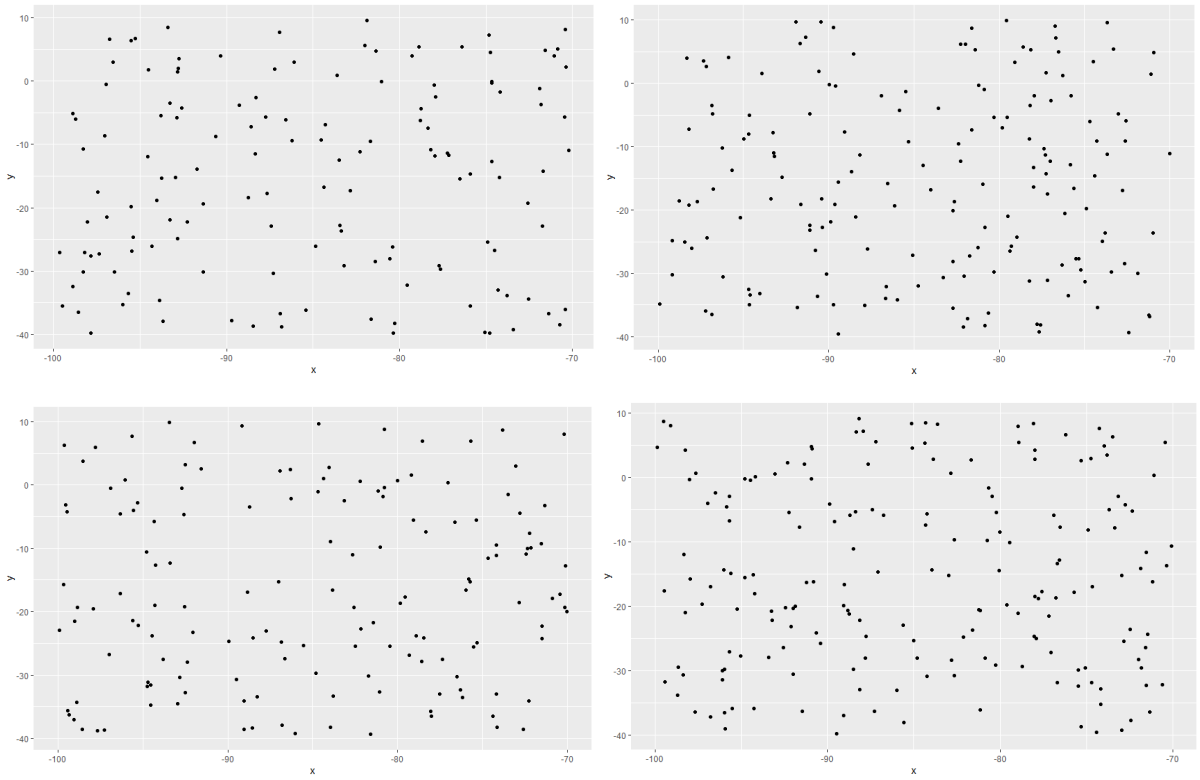

Figure 4. Results of running the random point distribution in R several times.

Running the R-script several times to generate random points in a defined bounding box shows that the results are different each time (see figure 4). Although the distribution of the dots appears similar at first glance, which could be related to the defined intensity, there are clear differences between the runs. This means that the result is probabilistic, since the outcome is random. With a deterministic result, the outcome is completely determined by the initial conditions, meaning all four figures should be the same. Accordingly, there are also many realizations of a spatial random distribution (H_0), since the results exhibit random variability and many potential outcomes are possible.

NNA as Nonparametric Test

Assignment 4

Carry out the same NNA Test on Westminster Crime as a nonparametric test.
• Nonparametric means we do not assume a distribution for test statistic z
• Instead, we calculate multiple spatial random distributions by iteration (sometimes called permutation test) and compare those theoretical distributions with the empirical distribution to get pseudo emp. p.

1. *Calculate empirical MNND (same procedure as with parametric test)*

```
##Load libraries##
library(sf)
library(spatstat)

##1. Calculate empirical MNND (same procedure as with parametric test)##
# Load shapefile as sf Objekt
west.admin<- sf::st_read("WestminsterCrime/WestminsterAdmin-sub.shp")
crimeWestminster<- sf::st_read("WestminsterCrime/WestCrime.shp")

plot(west.admin)
plot(crimeWestminster)

#convert sf object to ppp object
crimeWestminster.ppp <- spatstat.geom::as.ppp(crimeWestminster)

# calculate empirical mean nearest neighbor distance
nnd <- spatstat.geom::nndist(crimeWestminster.ppp)
mnnd_emp <- mean(nnd)
```

Codeblock 1. Calculating the empirical MNND.

Result: 50.88

2. *Calculate theoretical distributions by iteration*

```
##2.  Calculate theoretical distributions by iteration##
#generate 100 random point distributions and calculate theo. MNND for each
mnnd_theo <- c()
for (i in 1:100){
  west.admin.sample <- sf::st_sample(west.admin, nrow(crimeWestminster))
  west.admin.ppp <- spatstat.geom::as.ppp(west.admin.sample)
  nnd.sample<- spatstat.geom::nndist(west.admin.ppp)
  mnnd_theo <- c(mnnd_theo, mean(nnd.sample))
}
mnnd_theo
```

Codeblock 2. Calculating the theoretical distributions by generating 100 random point distributions.

3. *Get count theoretical MNND < emp. MNND*

```
##3.  Get count theoretical MNND < emp. MNND##
sum(mnnd_theo < mnnd_emp)
```

Codeblock 3. Calculating the number for which the theoretical MNND is smaller than empirical MNND.

Result: 100

4. *Calculate pseudo emp. p and interpret evaluate hypothesis*

We can calculate the pseudo empirical p as 1.0 because 100 out of 100 theoretical MNNDs are smaller than the empirical MNND, which means that 100% of the results are more extreme or more concentrated. Due to the two-sided hypothesis, we would normally multiply by 2, but since the value is already 1.0, it remains the same. Since the pseudo empirical p-value of 1.0 is significantly higher than the previously defined significance level of 0.01, we remain with the null hypothesis $H_{0,}$ which states that the spatial distribution of crime indices in the London ward of Westminster is random. As we cannot accept H_1, there is no evidence of a significant clustered or regular distribution.

Regression Analysis

Assignment 5

Investigate the effect of the explanatory variable Median Age 2013 (MdA_2013) on the dependent variable Average GCSE Score (AGc_2) by means of a linear regression model.

1. *Download LondonWardStats.zip from Blackboard*
2. *Implement your model in RStudio (see code above)*

The model was implemented using the proposed R Code by explaining the explanatory variable of unauthorized absence in 2013 with median age in 2013 (see R script). This leads to the following null and alternative hypothesis:

H_0: Median age in 2013 has no effect on student performance (GCSE).

H_A: Median age in 2013 has a significant effect on student performance (GCSE).

First, we needed to determine whether a linear model was appropriate by plotting the relationship between median age 2013 and average GCSE scores. This plot helps to visually inspect if there is a linear relationship and is also an indication of homoscedasticity. Since the spread of data points is relatively uniform across the range of the predictor variable, we can assume homoscedasticity and proceed with our linear regression.

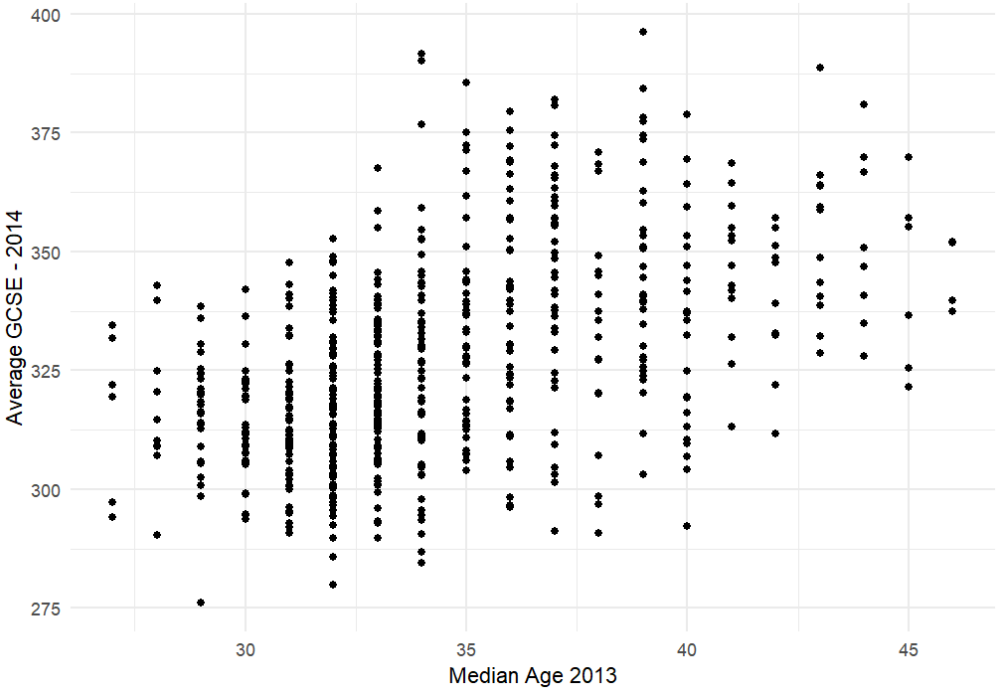

Figure 1. Relationship between Median Age 2013 and Average GCSE score 2014 to assess linearity and homoscedasticity.

Afterwards, the best-fitting linear regression line was aligned with the plot, which provides initial insights into the intercept and slope. In this example, a positive slope was found, which indicates that as the independent variable of median age 2013 increases, the dependent variable of average GCSE 2014 also increases.

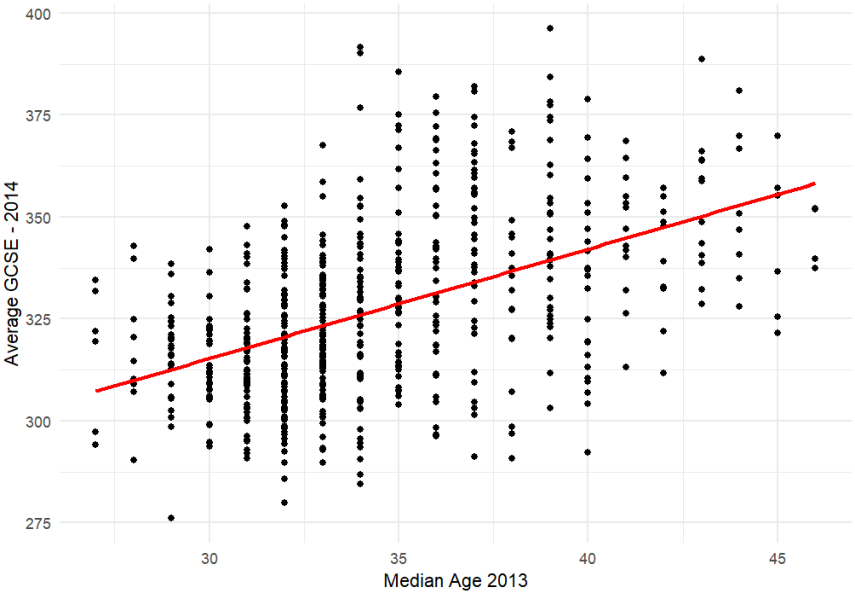

Figure 2. Fitted linear regression line showing that higher median age 2013 is associated with higher average GCSE score in 2014.

Additionally, the residuals were analyzed, since these should be normally distributed and not skewed. In this example, an approximately normal distribution centered around zero was found, which indicates a normal distribution of the residuals.

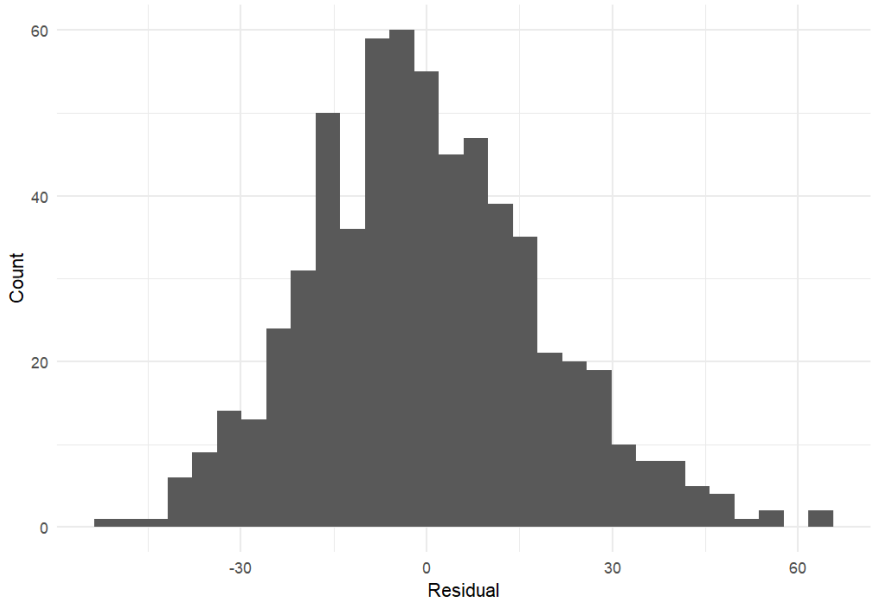

Figure 3. Histogram showing the distribution of the residuals of the linear regression model.

3. *Interpret outputs of the regression model in your own words.*

The results of the regression model can be best assessed using the model summary. The model summary shows various values which can be interpreted accordingly. First, the R-squared of 0.2419 indicates that approximately 24.2 % of the variance in average GCSE score can be explained by the median age 2013. In this example, the adjusted R-squared is quite similar as only one predictor was used here. Both the significance codes and the low p-value here indicate that the result is highly statistically significant. Since the p-value is below the previously defined significance level of 0.01, we must reject the null hypothesis and accept the alternative hypothesis, which states that median age in 2013 has a significant effect on GCSE score in 2014. Additionally, the model summary allows us to create the regression equation using the two coefficients intercept and slope:

$y = 234.6639 + 2.6867 * x$

Here y is the predicted average GCSE score and x is the median age in 2013.

```
Residuals:
    Min      1Q  Median      3Q     Max
-49.834 -12.915  -1.517  11.265  65.587

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 234.6639     6.6078   35.51   <2e-16 ***
MdA_2013      2.6867     0.1904   14.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.8 on 624 degrees of freedom
Multiple R-squared:  0.2419,    Adjusted R-squared:  0.2407
F-statistic: 199.2 on 1 and 624 DF,  p-value: < 2.2e-16
```

Figure 4. Summary of the linear model between the median age 2013 and the GCSE score 2014.

4. *Find a causal explanation for the observed correlation. Does your regression model show causality? What other variables could influence the dependent variable. Explain in a few sentences.*

While our model found a positive correlation between median age 2013 and GCSE score 2014, suggesting that an increase in median age 2013 tends to lead to a higher GCSE score 2014, this does not necessarily equate to causality. While causality implies that A causes B, correlation simply means that A and B tend to be observed at the same time. For this example, this means that median age 2013 and GCSE score 2014 were observed at the same time and there is a possibility that this also indicates causality, but it is important to understand that there is no actual proof. To identify underlying causes can be difficult since there is often a whole underlying causal network with different independent variables. In this example, one could also argue with the R^2 value of 0.2419, which indicates that over 75 % of the variance in average GCSE scores 2014 cannot be explained by this linear model. This clearly suggests that other variables are influencing the dependent variable, which in this example could be unauthorized absence in 2013 or even median house price in 2014 as an indicator of student socioeconomic status. To establish causality, more complex studies are required, in which a multilinear regression including several predictors could be a starting point.

Spatial Regression Analysis

Assignment 6

In Assignment 5 you have investigated the effect of the explanatory variable Median Age 2013 (MdA_2013) on the dependent variable Average GCSE Score (AGc_2).

1. *Include at least one more predictor / independent variable in your model that improves the overall model fit*

To improve the overall model fit, I included three predictors, namely Unauthorized Absence 2013 (UAAS_), Median Age 2013 (MdA_2013) and Median House Price 2014 (MHP__). The improved model has a R-squared of 0.5157, which means that around 51.6 % of the variance in average GCSE score can be explained by the defined predictors. The adjusted R-squared is slightly lower since it considers all three predictors. Both the significance codes and the low p-value here indicate that the result is highly statistically significant. Since the p-value is below the previously defined significance level of 0.01, we reject the null hypothesis and accept the alternative hypothesis, which states that the three defined predictors have a significant impact on GCSE score in 2014. Additionally, the model summary allows us to create the regression equation using the intercept and the coefficients of the predictors:

$y = x1 * -28.9911 + x2 * 1.1602 + x3 * 13.0217 + 150.7472$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 150.7472    20.9946   7.180 1.99e-12 ***
UAAS_       -28.9911     2.1690 -13.366  < 2e-16 ***
MdA_2013      1.1602     0.1797   6.457 2.15e-10 ***
log(MHP__)   13.0217     1.4573   8.935  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.05 on 622 degrees of freedom
Multiple R-squared:  0.5157,    Adjusted R-squared:  0.5134
F-statistic: 220.8 on 3 and 622 DF,  p-value: < 2.2e-16
```

Figure 1. Summary of the multilinear model between the three predictors and the GCSE score 2014.

2. *Visually inspect and interpret linearity of relationships between predictors and criterion variable as well as homoscedasticity. Transform variables if needed.*

To inspect and interpret the linearity of the relationships and assess the homoscedasticity between the predictors and the criterion variable, the relevant predictors were plotted in relation to the average GCSE score in 2014. For the first two predictors, Unauthorized Absence in 2013 and Median Age 2013, a relatively uniform spread of data points can be seen, indicating homoscedasticity. Here, a negative association between Unauthorized Absence 2013 and Average GCSE Score 2014 and a positive association between Median Age in 2013 and Average GCSE Score 2014 can be seen.

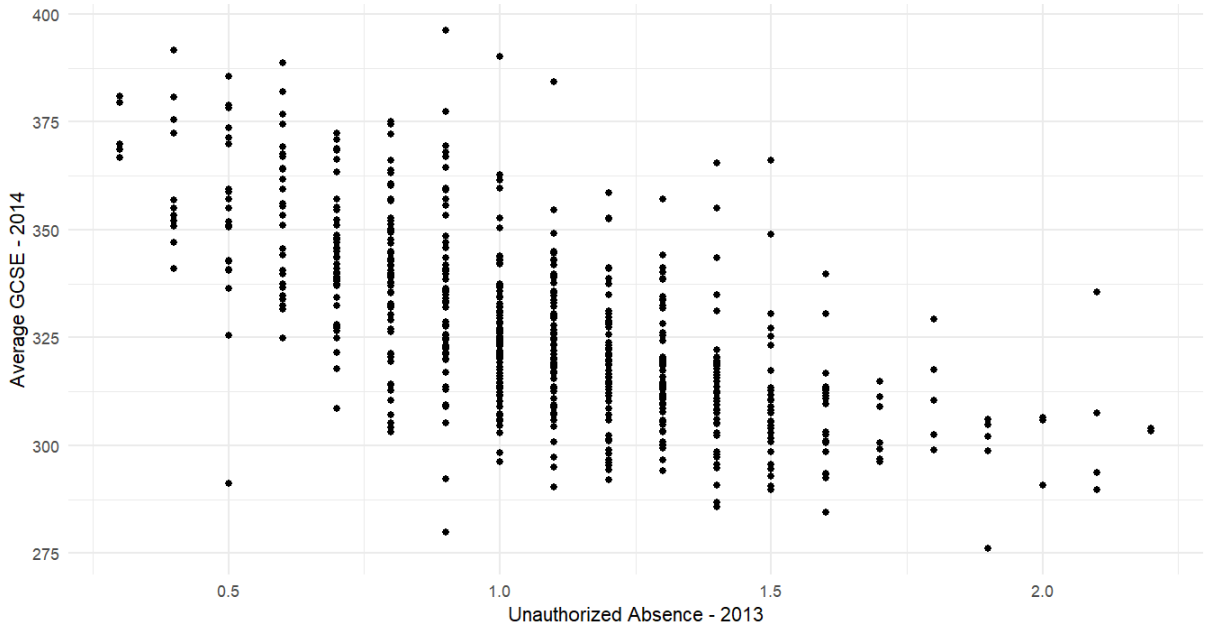

Figure 2. Relationship between Unauthorized Absence 2013 and Average GCSE score 2014.

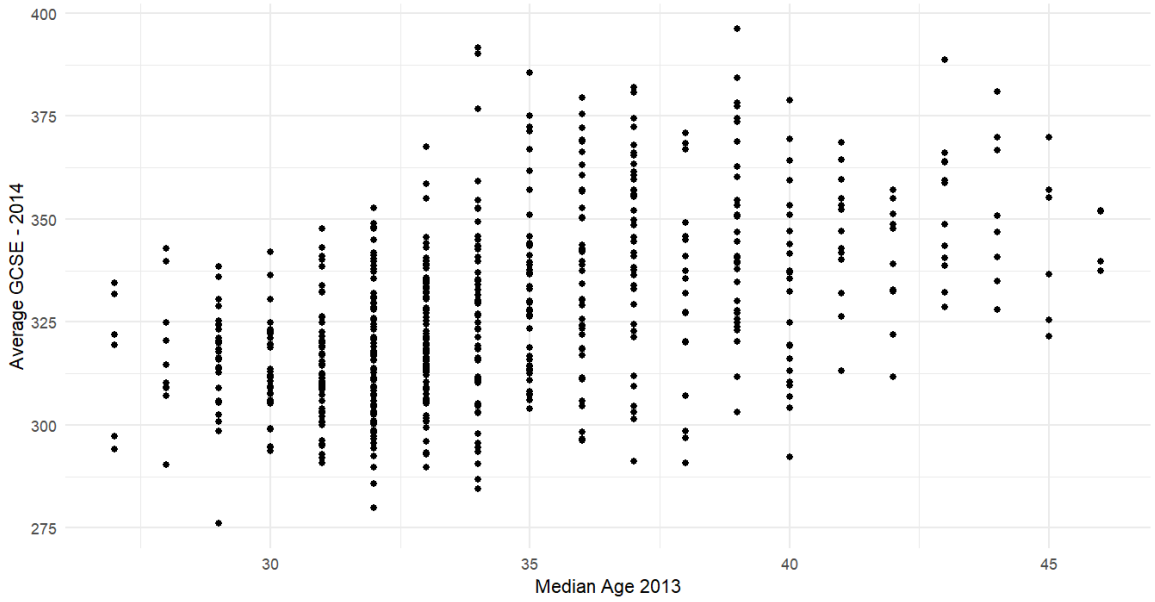

Figure 3. Relationship between Median Age 2013 and Average GCSE score 2014.

For the third predictor Median House Price 2014, the initial plot does not show linearity, but indicates that the data points are right skewed. At the same time, the data points do not show homoscedasticity, but heteroscedasticity.

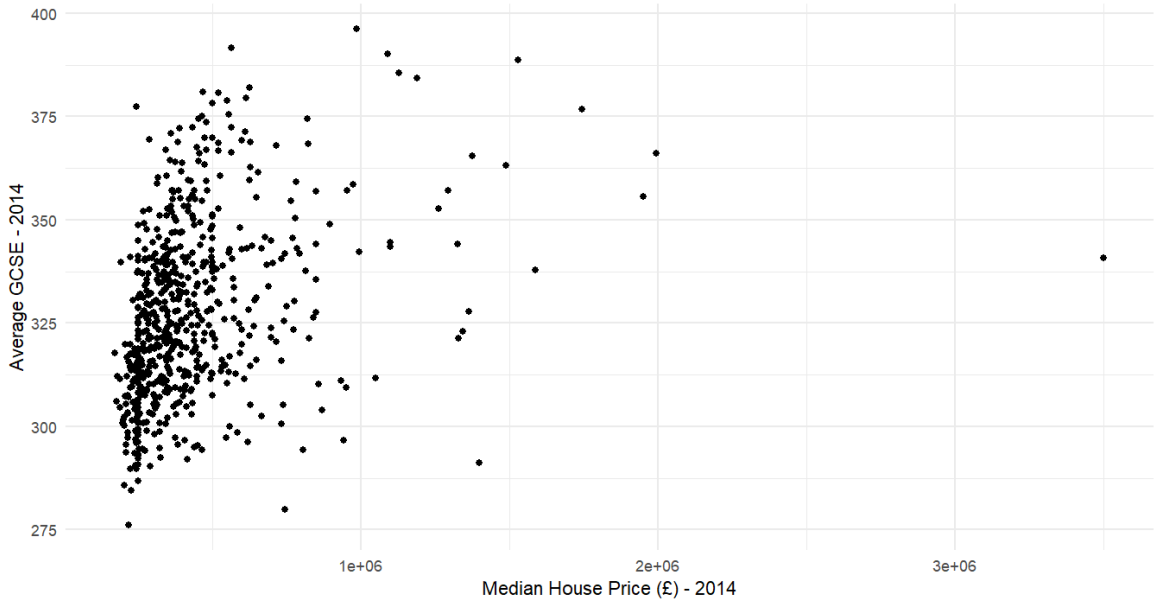

Figure 4. Relationship between Median House Price 2014 and Average GCSE score 2014.

To be able to integrate this predictor into multilinear regression nevertheless, the data was transformed with the logarithm. The result appears to be more linear than before the transformation, even if it is still slightly skewed. At the same time, perfect homoscedasticity can't be determined. Nevertheless, this transformed data can now be used for multilinear regression.

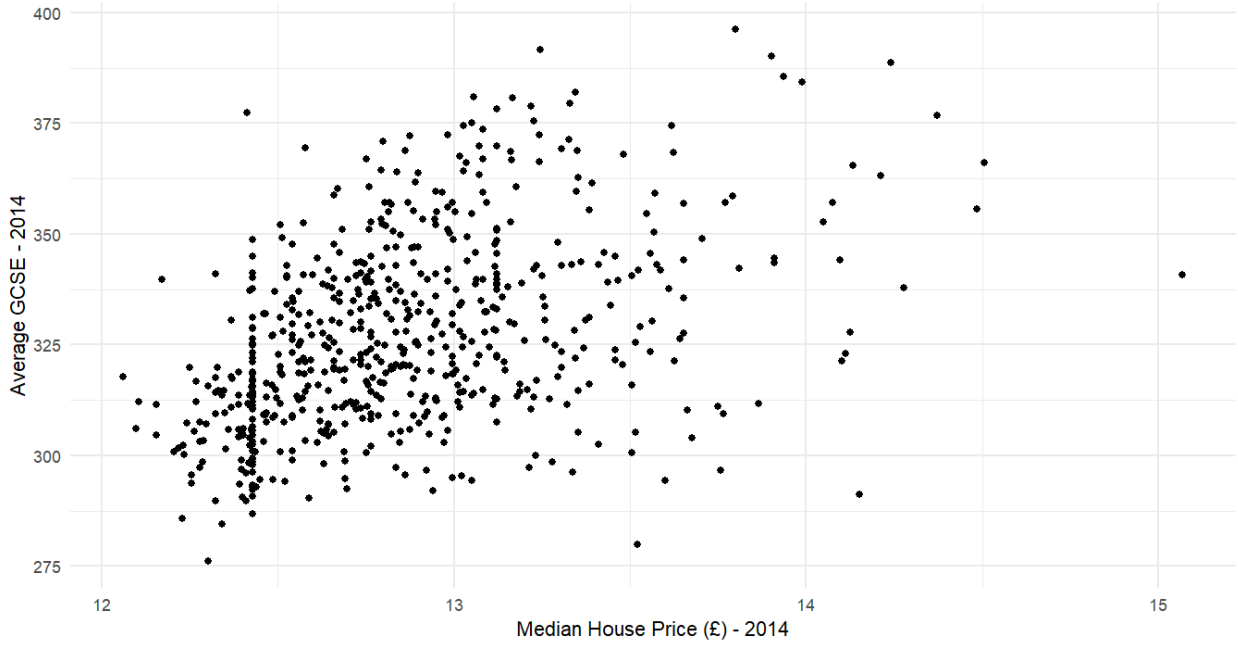

Figure 5. Relationship between transformed (log) Median House Price 2014 and Average GCSE score 2014.

3. *Test for multicollinearity among predictors*

To test for multicollinearity among predictors, three Pearson tests were conducted. The first test for multicollinearity between Unauthorized absence 2013 and transformed Median House Price 2014 shows a correlation of -0.31. At the same time, this results in a very low p-value, which can be interpreted as highly significant. If we assume that the rule of thumb states that Pearson correlation above 0.9 indicates that two independent variables are highly correlated and collinear, this is not the case for these two predictors, as there is a low correlation (StrataScratch, 2025).

Similarly, the test for multicollinearity between Unauthorized absence 2013 and Median Age 2013 shows a very low p-value, but at the same time the correlation of -0.53 is below the threshold, meaning that the two predictors are not multicollinear. Finally, multicollinearity was tested between the transformed Median House Price 2014 and the Median Age 2013, again with a low p-value and a correlation that does not indicate multicollinearity. In summary, this means that all three predictors can be used for multilinear regression, since no problematic multicollinearity could be identified.

4. *Map and visually interpret spatial autocorrelation of your model's residuals*

After the normal residuals were inspected, the clustered residuals were mapped and visually examined. The visual interpretation shows both a local positive spatial autocorrelation in some regions where clusters of similar residual values were found and a local negative spatial autocorrelation in regions where neighboring areas with opposite residual values were identified. This shows that there is definitely no random distribution of the inspected data.

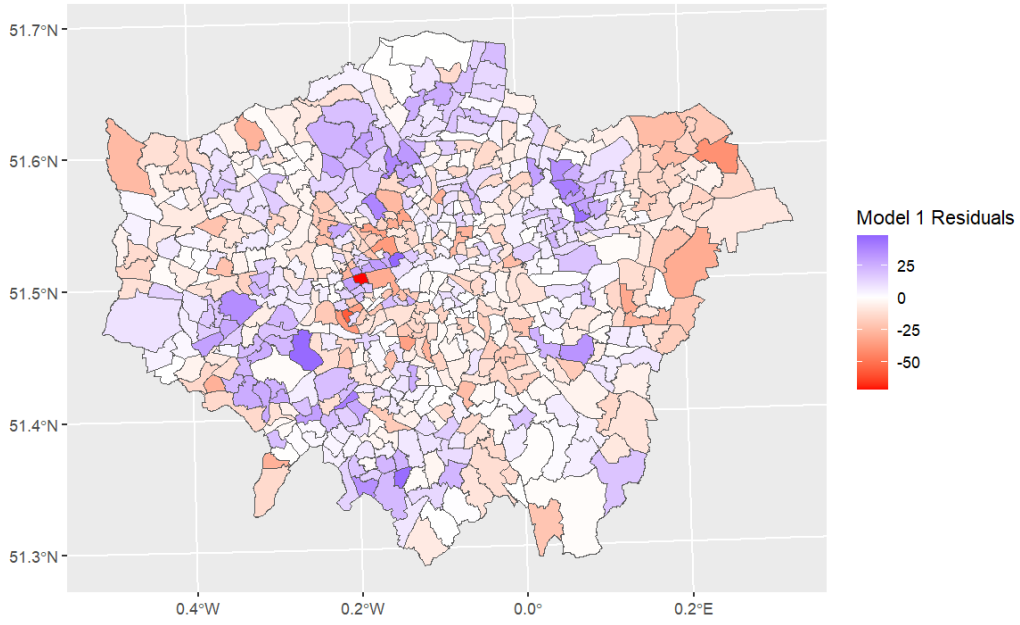

Figure 6. Spatial distribution of the residuals of a multilinear regression.

To calculate the statistical spatial autocorrelation, Moran's I was also applied. First, a very low p-value was determined, which shows that the observed spatial autocorrelation is highly statistically significant.

At the same time, the value of the spatial autocorrelation is 0.271, which indicates a moderate positive spatial autocorrelation in which similar residuals accumulate

5. *If your residuals are spatially autocorrelated, conduct a Spatial Lag Regression. How does that change the coefficients of your model?*

Since the mapping of residuals and Moran's I showed a strong spatial autocorrelation, a Spatial Lag Regression was performed using the NN matrix with the 4 nearest neighbors. This Spatial Lag Model includes a spatially lagged dependent variable as an additional predictor. This gives the coefficients of the model lower values, which is why the initial regression equation must be adjusted accordingly. Specifically, the intercept changed from 150.7472 to 90.2, the UAAS_ coefficient from -28.9911 to -24.3, the MdA_2013 coefficient from 1.1602 to 0.817 and the transformed MHP__ from 13.0217 to 9.87.

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	rho	0.330	0.0419	7.88	3.33e-15
2	(Intercept)	90.2	20.9	4.32	1.53e- 5
3	UAAS_	-24.3	2.13	-11.4	0
4	MdA_2013	0.817	0.175	4.67	3.01e- 6
5	log(MHP__)	9.87	1.45	6.81	9.86e-12

Figure 7. Summary of the conducted Spatial Lag Regression.

6. *Upload zip-file (assign6_name.zip) that contains your RScript (no files!) + PDF File with plots and interpretations.*

References

StrataScratch. (2025). *A beginner's guide to collinearity: What it is and how it affects our regression model*. Retrieved May 17, 2025, from https://www.stratascratch.com/blog/a-beginner-s-guide-to-collinearity-what-it-is-and-how-it-affects-our-regression-model/

Global Spatial Autocorrelation

Assignment 7

1. *Classify the crater dataset into 5 size groups (attribute: DIAM_C_IM)*

To classify the crater dataset into 5 size groups, I used the quantile method on the symbology tab in ArcGIS Pro. This method was chosen because it distributes the observations equally across the class interval, creating unequal class widths but the same frequency of observations per class. In a second

step, I was then able to select the attributes according to the corresponding diameters in kilometers from a circle fit (DIAM_C_IM) for each class.

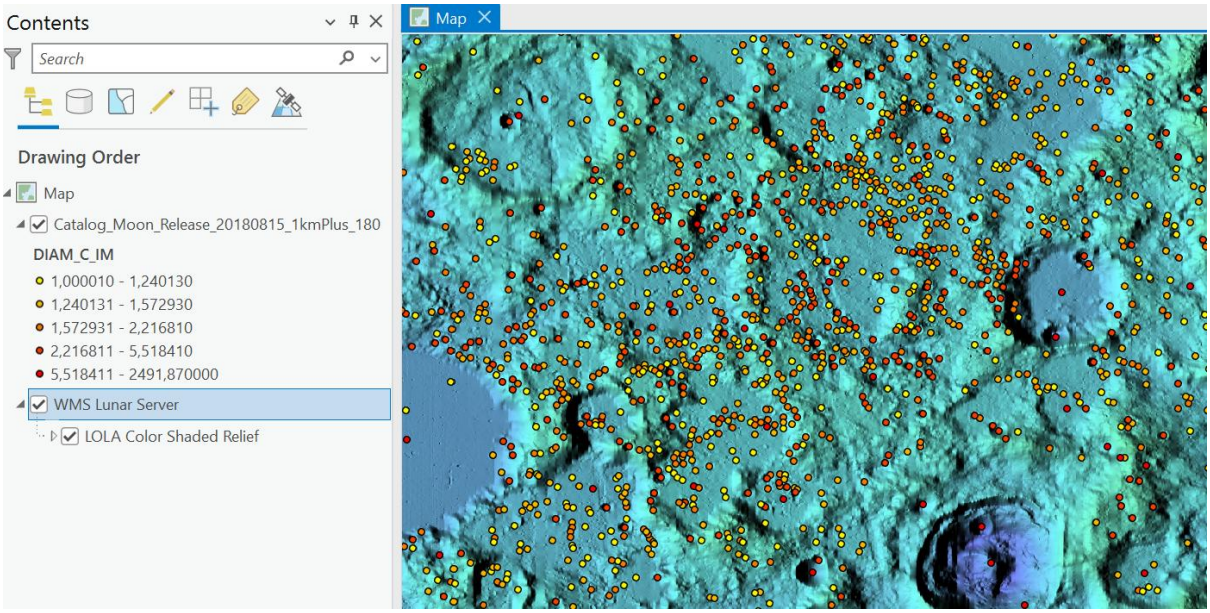

Figure 1. Classified crater dataset using the quantile method in ArcGIS Pro.

2. *Calculate Global Spatial Autocorrelation of crater orientation (D_E_ANG_IM) for each class individually (use inverse distance as Conceptualization of Spatial Relationship).*

I calculated the Global Spatial Autocorrelation (Global Moran's I) for each category using the appropriate geoprocessing tool in ArcGIS Pro. For each class, the attributes that are part of that class were pre-selected using the *Select By Attributes* tool, as mentioned earlier. For each run, the conceptualization of the spatial relationships was defined as Inverse distance. Furthermore, a report was created for all five categories.

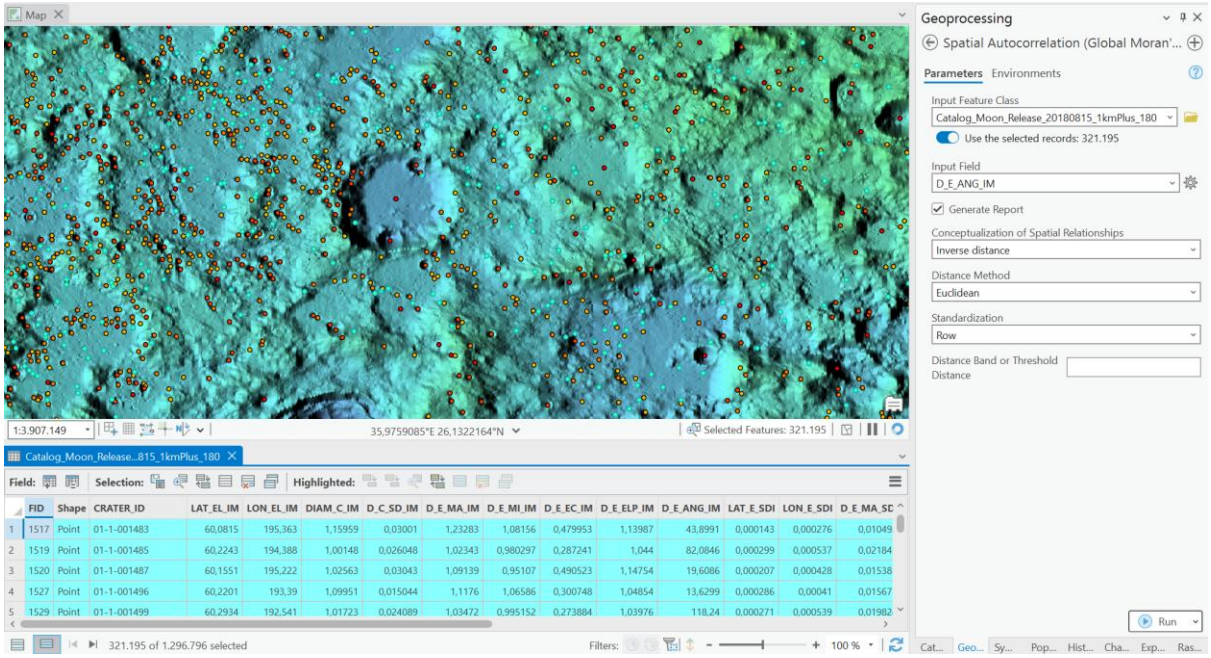

Figure 2. Calculating the Spatial Autocorrelation for the first category with ArcGIS Pro.

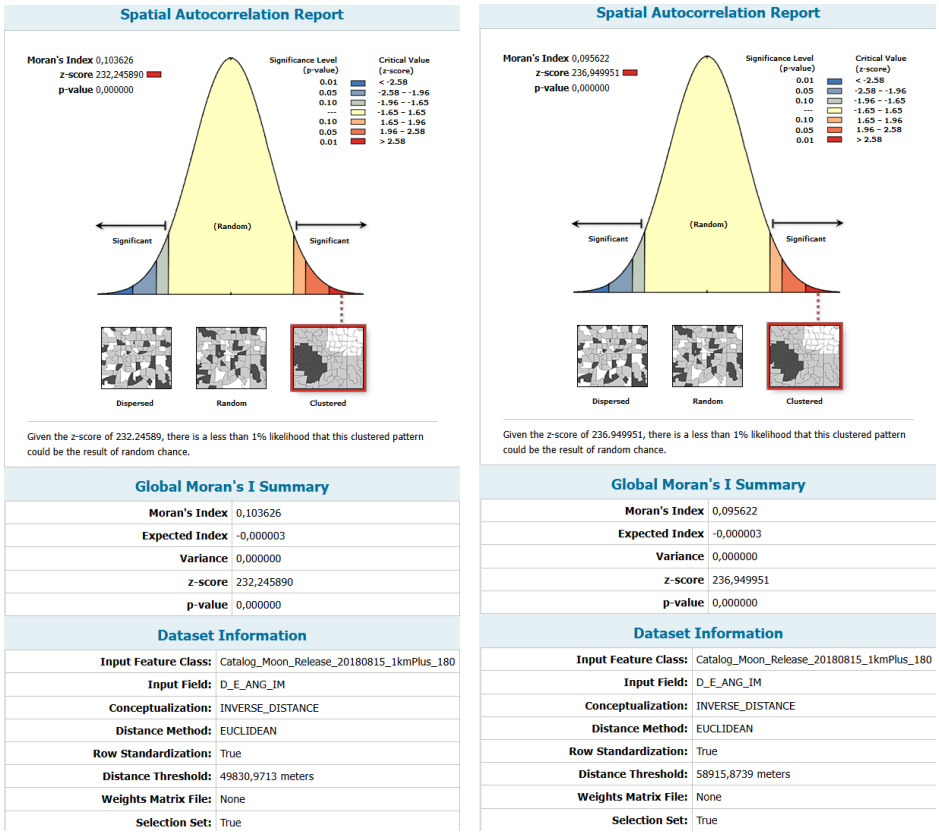

Figure 3. Spatial Autocorrelation Report for the first and second category using ArcGIS Pro.

3. *Make conclusions about the assumed origin of respective classes of craters (primary or secondary craters).*

For each of the five size groups, we define a significance level of 0.01 and a null-hypothesis H_0, which states that the orientation of the craters in the corresponding size range is randomly distributed. Furthermore, based on the analysis already performed, we expect the orientation of the primary, large craters to be randomly distributed, while we assume a spatial autocorrelation for the orientation of the secondary, small craters.

The Spatial Autocorrelation (Global Moran's I) of the first class with the smallest diameter in kilometers from a circle fit (DIAM_C_IM) has the highest Moran's Index of all five classes at around 0,104. This is an indication of a positive spatial autocorrelation. At the same time, we have a very high z-score of about 232,246, indicating that the observed pattern is most likely not random. At the same time, the report gave a p-value of 0, confirming that the pattern is statistically significant, which is why we have to reject the H_0 hypothesis.

Although Moran's Index and Z-Score values decrease with the increase in diameter in kilometers from a circle fit (DIAM_C_IM), the following four categories all show a weak to moderate positive autocorrelation with high z-scores, indicating that the categorized craters are unlikely to be random. Simultaneously, the p-value for all categorized size classes is 0, showing a statistically significant pattern. In this context, the large sample size for all five classes should be mentioned, as this can lead to a smaller standard error, which affects the statistical inference of Global Moran's I.

All in all, it can be said that the first four classes confirm the assumption that secondary craters are not randomly distributed, but show a significant spatial autocorrelation. The results of the last category with a Moran's Index showing a weak positive autocorrelation and a comparatively low z-score of about 19,16 indicate that it could be a mixed class of primary and secondary craters. To achieve a better distinction between primary and secondary craters and their spatial randomness or autocorrelation, other class intervals might have been more suitable.

Table 1. Results of the Spatial Autocorrelation Report for the five classified size groups.

Class	Class Interval	Sample Size	Moran's Index	Z-Score
1	1,000010-1,240130	321.195	0,103626	232,245890
2	1,240131-1,572930	303.596	0,095622	236,949951
3	1,572931-2,216810	303.288	0,073066	290,634860
4	2,216811-5,518410	299.088	0,040854	189,619104
5	5,518410-2491,870000	69.604	0,013798	19,160875

Local Spatial Autocorrelation

Assignment 8

1. *Select a study area (one primary many secondary craters).*

I tried to select a study area with one large primary crater and many secondary craters. To do this, I used the *Select by Rectangle* tool to pick the attributes for my study area. As a result, 2372 attributes were selected and used for calculating the Local Spatial Autocorrelation of crater orientation.

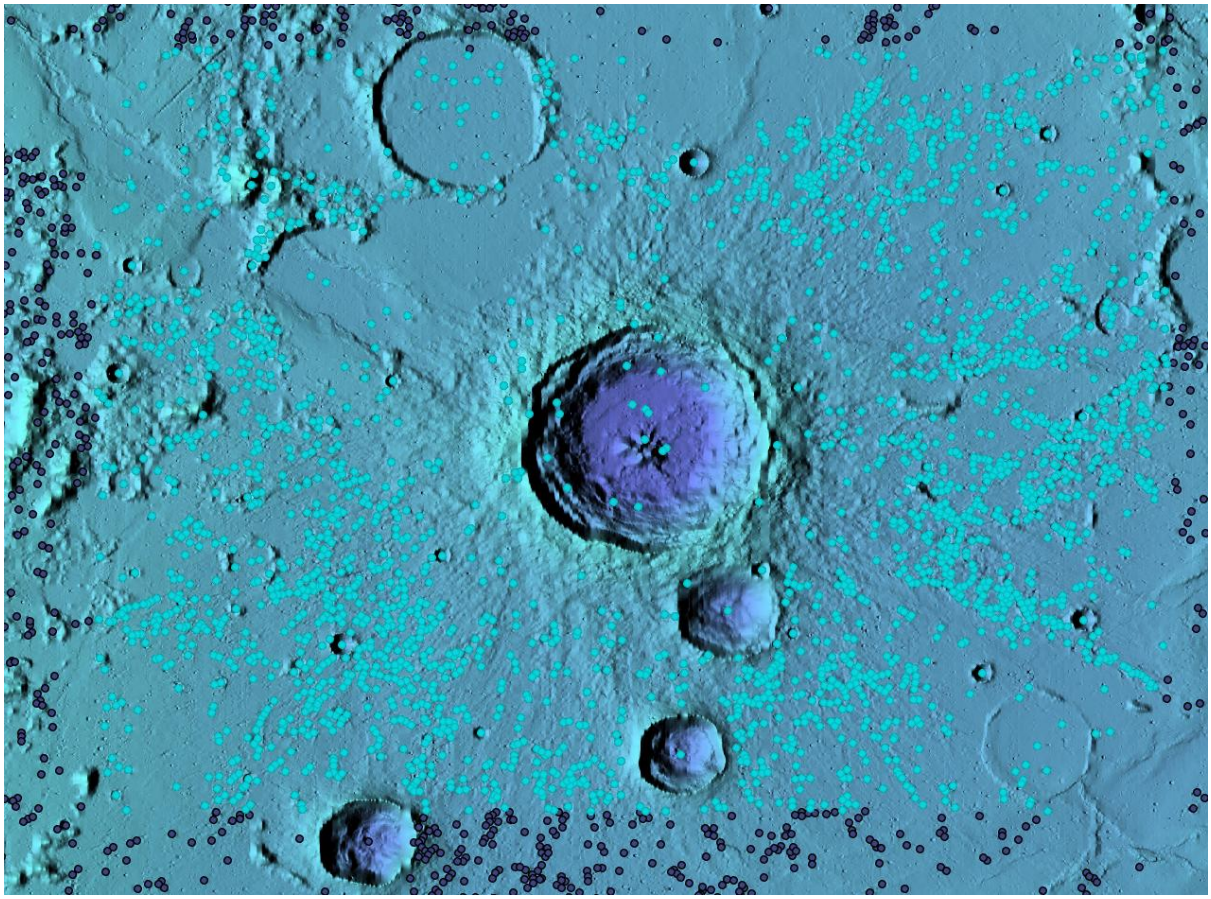

Figure 1. Study area showing the primary crater and all 2372 selected attributes in cyan.

2. *Calculate Local Spatial Autocorrelation of crater orientation (D_E_ANG_IM) for the selected study area (use inverse distance as Conceptualization of Spatial Relationship and FDR correction).*

I calculated the Local Spatial Autocorrelation of crater orientation for the study area using the *Cluster and Outlier Analysis (Anselin Local Moron's I)*. Thereby, I used *Inverse Distance* as *Conceptualization of Spatial Relationships* and also applied the *False Discovery Rate (FDR) Correction*.

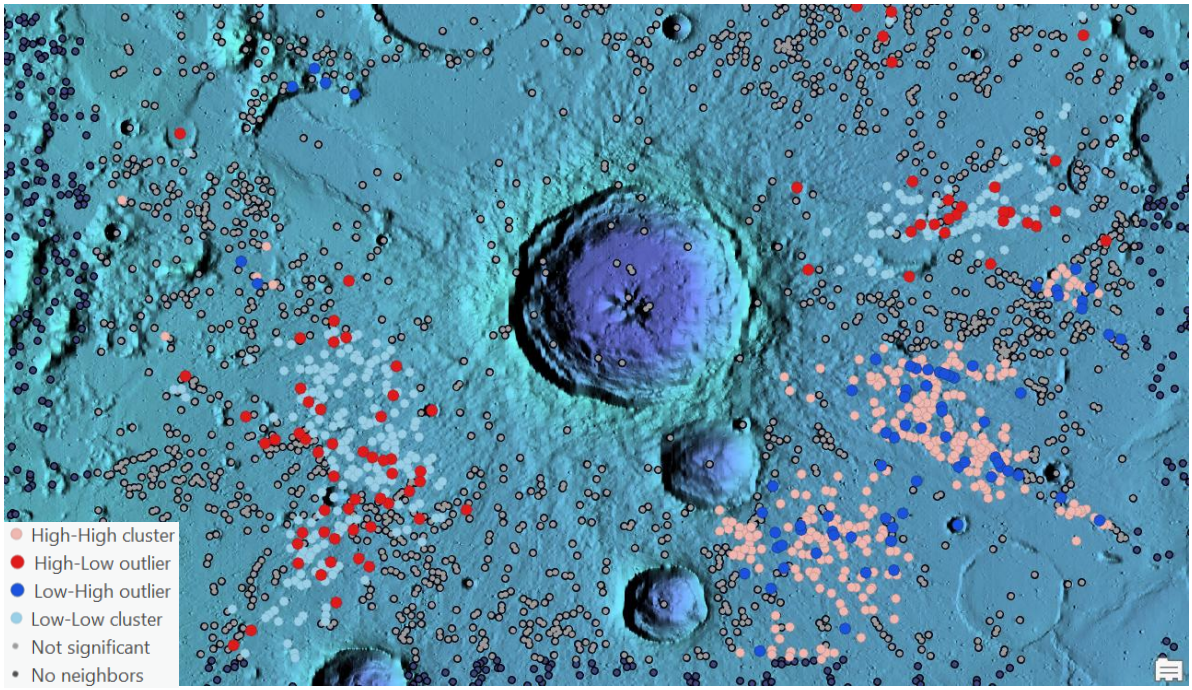

Figure 2. Local Spatial Autocorrelation of crater orientation for the study area.

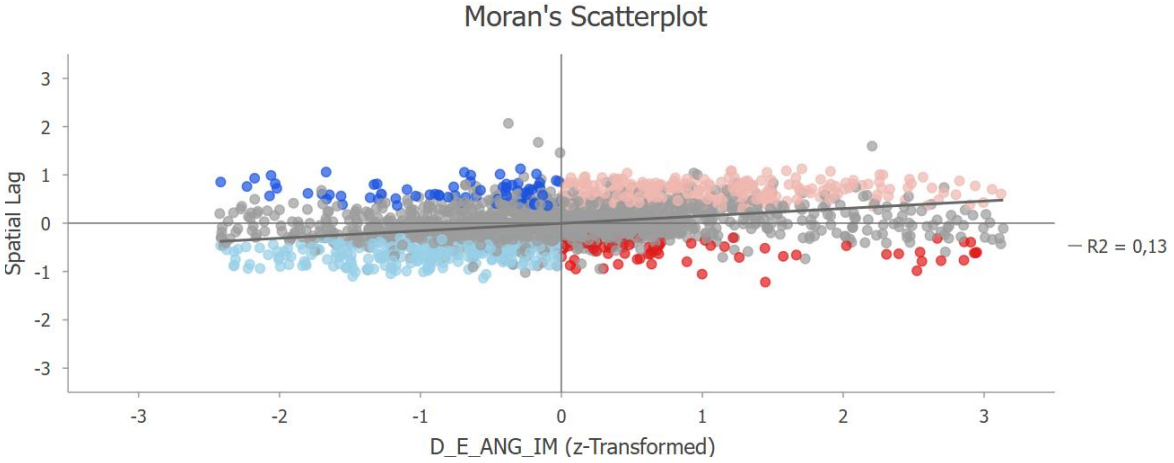

Figure 3. Scatterplot of the Local Spatial Autocorrelation of crater orientation for the study area.

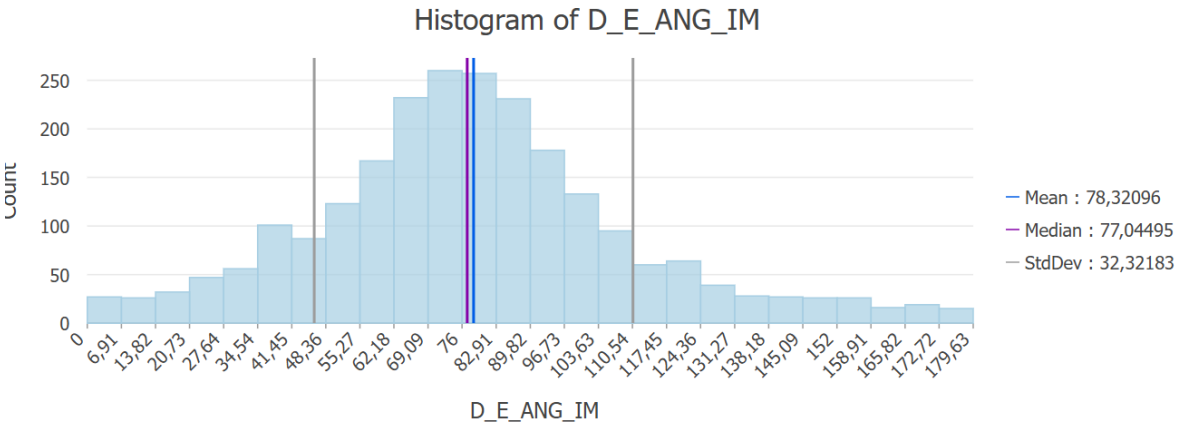

Figure 4. Histogram of the Local Spatial Autocorrelation of crater orientation for the study area.

3. *Which craters presumably stem from materials that were ejected due to the primary impact (secondary cratering)?*

Both the high-high and low-low clusters indicate that the craters in these areas probably stem from the same primary impact. Thereby, high-high clusters suggest craters with east-west orientation (positive z-scores) surrounded by similarly oriented neighbors, while low-low clusters indicate craters with west-east orientation (negative z-scores) surrounded by similarly oriented neighbors. On the other hand, the low-high and high-low outliers show areas where neighboring craters have a different orientation. These outliers are therefore not related to the impact of the analyzed primary impact, but could rather point to impacts from other events.

4. *What is FDR correction? Why do we use FDR correction in the case of local autocorrelation analysis?*

The False Discovery Rate (FDR) Correction is a statistical method that is used to control the expected proportion of incorrectly rejected null hypotheses (false positives). This is very important in the case of local autocorrelation, as this analysis considers each feature within the context of neighboring features, which leads to issues with multiple testing and spatial dependency. This might increase the probability of Type I errors.

Given a significance level of 5% and a dataset of 10,000 features, we would expect 500 incorrect results where we would get statistically significant p-values, even though the underlying pattern is random. At the same time, according to the first law of Geography, features that are close to each other tend to be similar, which can lead to artificially inflated statistical significance when neighboring features are considered.

Therefore, the FDR correction reduces the critical p-value thresholds according to the number of input features and the neighborhood structure employed. In this context, FDR correction estimates the number of false positives for a given confidence level and adjusts the critical p-value accordingly, thus reducing the likelihood of misidentification of spatial clusters (Esri 2025).

5. *Select and export craters to a new layer that show significant autocorrelation on a 1% significance level. The probability for what error type increases as the significance level is decreased? Explain in a few words why the probability for this error type increases.*

I selected craters that show a significant autocorrelation at a 1% significance level using the *Select By Attributes* tool and picking all craters that have a pseudo p-value of less than 0.01. I then exported the 617 selected craters as a new layer using the *Export Features* tool.

The selected features include both craters with positive and negative autocorrelation, since both can be significant. Here, the selected high-high and low-low clusters, which all have a pseudo p-value below 0.01, represent a positive local spatial autocorrelation with similar orientations. On the other hand, low-high and high-low outliers are also part of the new layer as they have negative local spatial autocorrelation, indicating that the features are significantly different (pseudo p-value < 0.01) from their neighbors. At this point, it should be mentioned that some clusters as well as outliers were not exported to the new layer since they had a pseudo p-value above 0.01.

By reducing the significance level to 0.01, the probability of Type II errors increases. This means that a real pattern could exist, but the statistical test may not recognize it. This was the case in this particular example, as some previously recognized clusters and outliers were not exported to the new layer. However, it should also be mentioned that with an increase of Type II errors, Type I errors, where craters are incorrectly identified as clusters or outliers, are reduced.

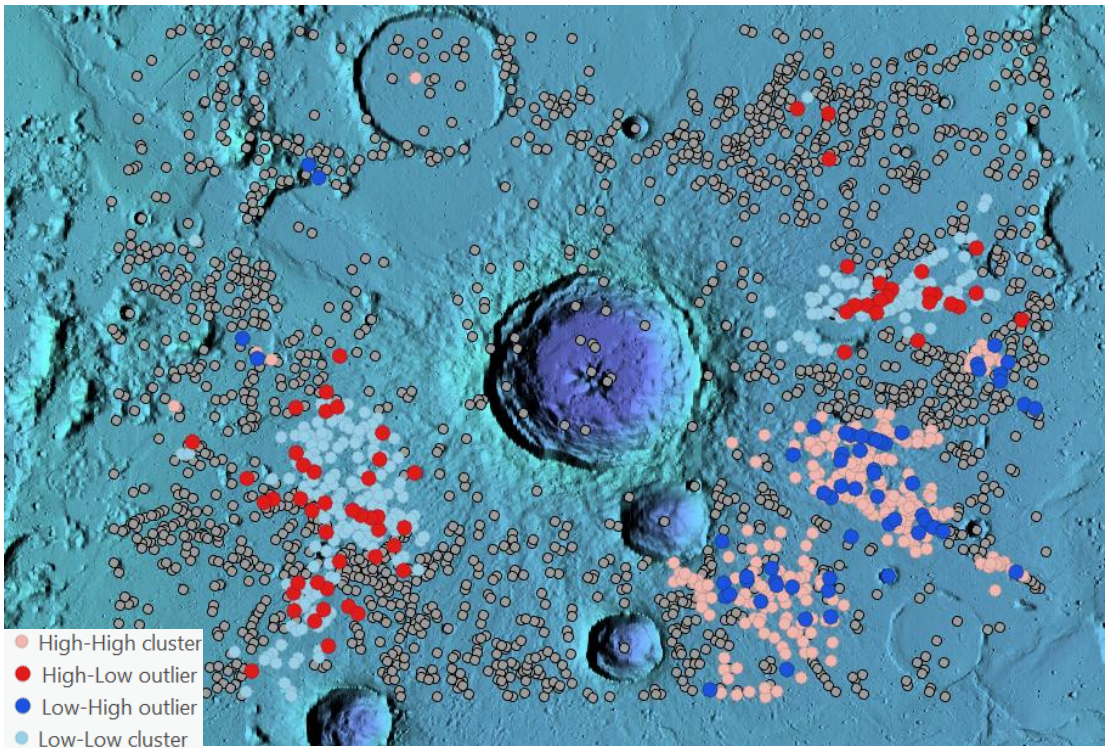

Figure 5. Craters showing a significant autocorrelation at a 1% significance level in blue and red. Craters above this level are shown in gray.

References

Esri. (2025). *What is a z-score? What is a p-value?—ArcGIS Pro documentation*. Retrieved June 5, 2025, from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm#ESRI_SECTION1_2C5DFC8106F84F988982CABAEDBF1440

Variography and Probabilistic Interpolation

Assignment 9

1. *Download Swiss Rainfall Data from BB, Attribute Z_1_10mm is yearly precipitation in cm*
2. *Visually inspect distribution of data by means of histogram or QQ-Plot, we assume stationarity*

The histogram of yearly precipitation in cm for Switzerland shows a right-skewed distribution with many rather low yearly precipitation values compared to the overall values. The mean value is around 192.21 cm of precipitation per year with a median of 164. At the same time, a relatively high standard deviation of around 114.35 cm can be seen, explained by the high yearly precipitation values on the far right, which may indicate outliers.

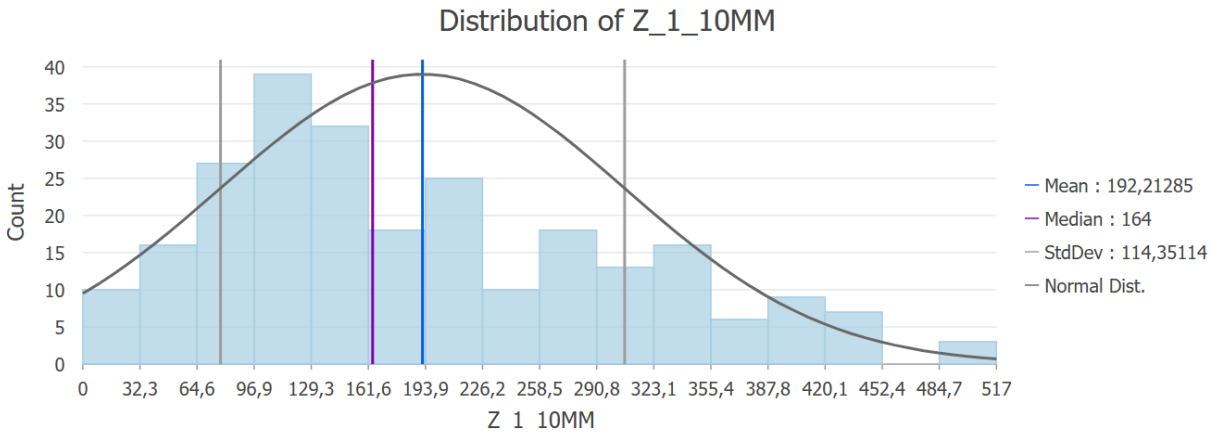

Figure 1. Histogram of the distribution of the yearly precipitation in cm (without transformation).

Similarly, the QQ plot, which compares the input data with a normal distribution, shows that the yearly participation in cm for Switzerland is not normally distributed. At the same time, the plot visualizes in particular the outliers with extremely low (0) and high yearly precipitation values.

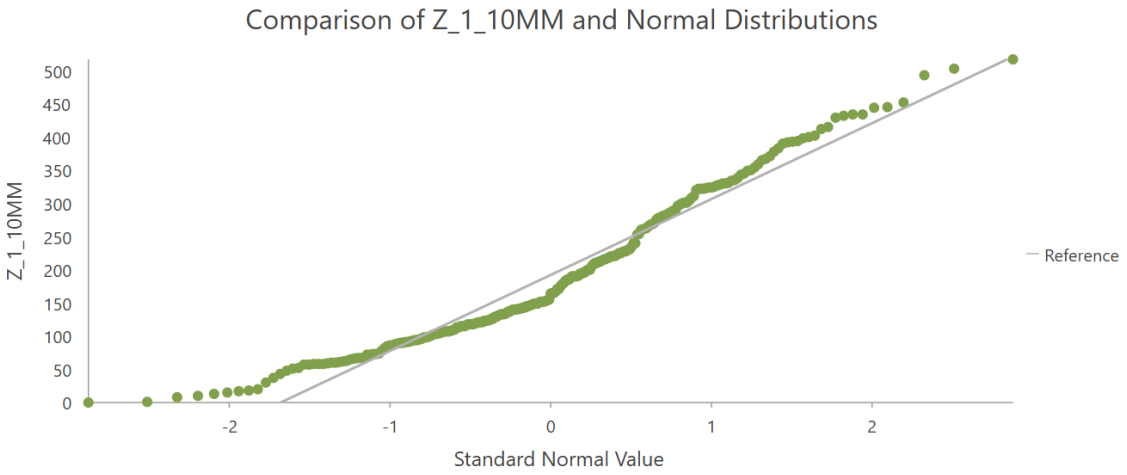

Figure 2. Histogram of the distribution of the yearly precipitation in cm (without transformation).

3. *To approximate a normal distribution, what transformation do you suggest (you may have to remove negative and zero values before)?*

Since both the histogram as well as the QQ plot showed that the data is right-skewed and therefore not normally distributed, a transformation technique can be applied. This reduces the skewness, minimizes outliers and finally approximates the normal distribution. A suitable approach in this example is a logarithmic transformation, whereby it should be noted that the previously determined 0 values must first be removed for calculating the logarithmic transformation.

The histogram of the transformed yearly precipitation is no longer right skewed and more closely approximates a normal distribution. This can also be seen in the mean and median values, which are now closer together, as well as in the comparatively low standard deviation of around 0,79. However, it should be noted that this transformation also generates some bins with low values on the left side.

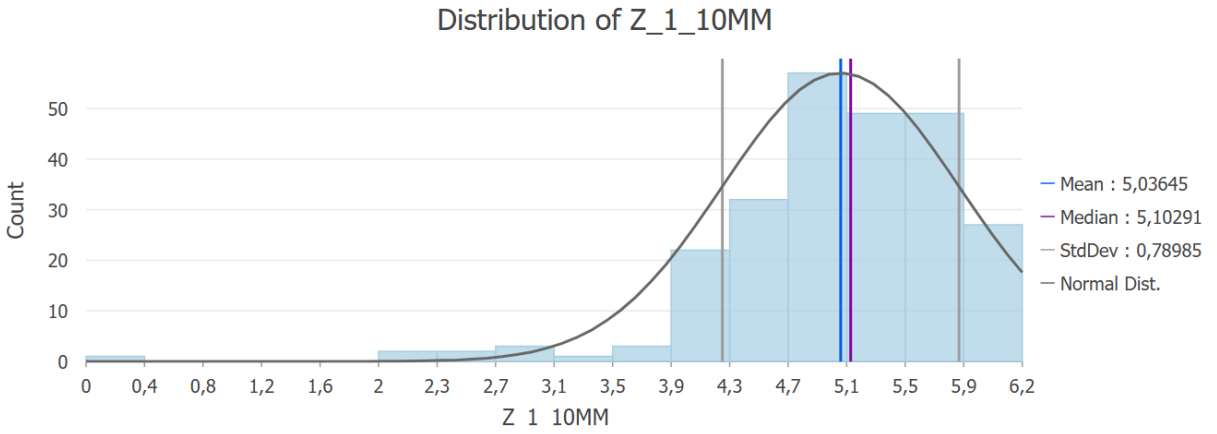

Figure 3. Histogram of the distribution of the yearly precipitation in cm (with logarithmic transformation).

The outliers with the small bin counts are also clearly visible in the QQ plot. At the same time, however, the middle part of the transformed distribution now approaches a normal distribution, which is illustrated by the green curve that now better depicts the diagonal of the normal distribution.

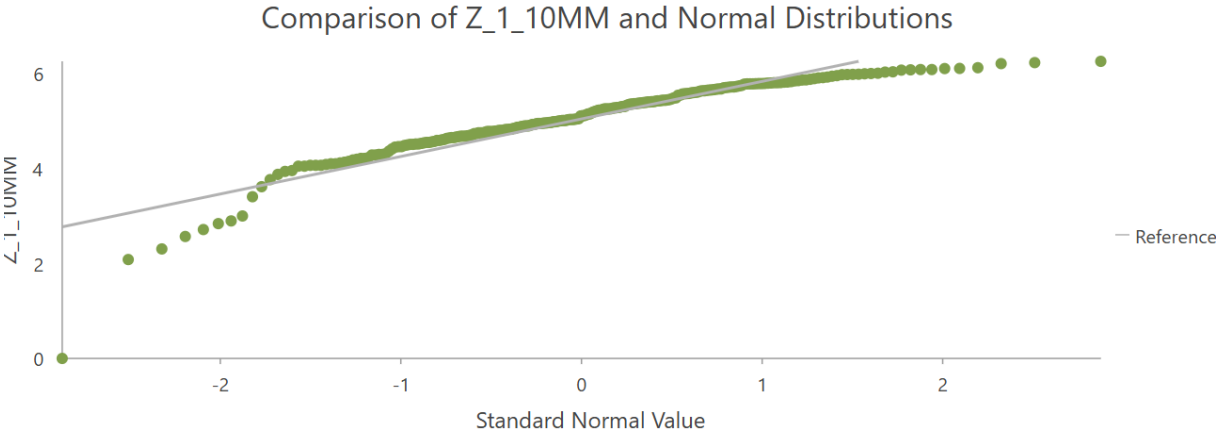

Figure 4. Histogram of the distribution of the yearly precipitation in cm (with logarithmic transformation).

4. *Open the Geostatistical Wizard, select Kriging/CoKriging and data field Z_1_10mm*
5. *In the next context menu select „Prediction" under Ordinary Kriging, chose an appropriate transformation, do not remove trends*

As already mentioned, a logarithmic transformation was chosen to approximate a normal distribution. At the same time, although the precipitation shows a trend based on topography, this trend was not removed accordingly, since the result should show both the overall trend and the differences or variations which are not based on the trend.

6. *Empirical Semivariogram: Chose lag size and number of lags.*

A lag size of 16500 with a total of 10 lags was selected for the empirical semivariogram. These settings come very close to the maximum lag distance between west and east divided by two, which is a standard approach in geostatistics.

7. *Select a theoretical semivariogram that approximates the empirical semivariogram. You may choose an anisotropic semivariogram given that semivariances are direction dependent.*

First, an isotropic semivariogram with a Gaussian function was selected, which was fitted to the empirical semivariogram. The semivariogram was then changed to an anisotropic semivariogram to consider both distance and direction.

8. *Interpret the semivariogram (empirical and theoretical with model parameters)*

In general, the semivariogram is represented by blue crosses showing the average semivariance of point pairs within a lag bin. The red dots show the raw semivariance values between individual pairs, which indicate the variability within the individual bins. For the isotropic semivariogram, the blue line represents the fitted Gaussian model. The model contains a low nugget, which represents an error of the theoretical semivariogram. The sill is the maximum semivariance and the partial sill is the portion of the total sill attributed to spatial structure. In this example, the range is approximately 82 km, which defines the spatial extent of the autocorrelation, whereby the values within this distance influence each other.

In addition to the graph of the semivariogram, a semivariogram map was created that visualizes the spatial structure of the semivariogram. The blue and green blocks in the center represent a lower semivariance and a higher spatial autocorrelation, while the orange and red blocks at the edges show a higher semivariance and a lower spatial autocorrelation.

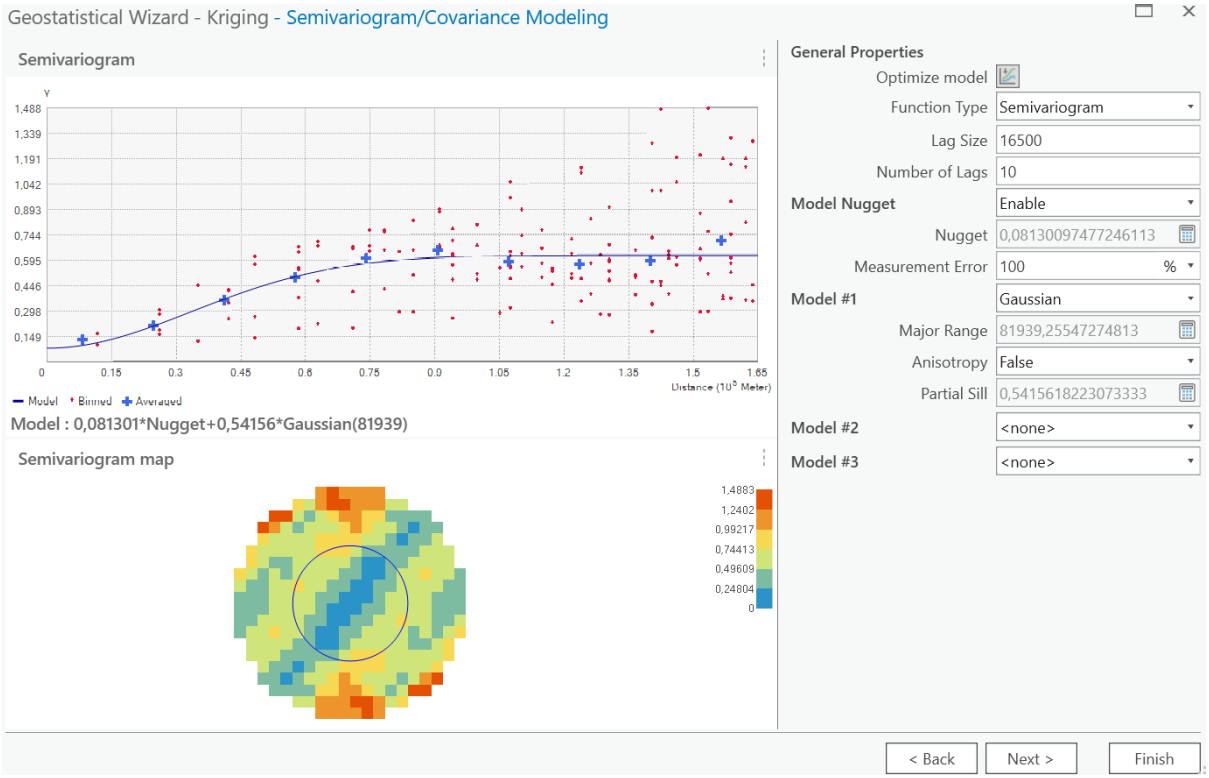

Figure 5. Isotropic semivariogram based on a Gaussian function for the study area.

Since the isotropic semivariogram ignores directional differences and is only valid for approximately symmetrical patterns, an anisotropic semivariogram was also selected, which has a directional ellipsoidal structure. In comparison to the isotropic semivariogram, the theoretical semivariogram, which is based on the Gaussian function, is now plotted for several directions to compare how the spatial dependence changes with the direction. Here, a nugget of around 0.04 can be seen, which represents the measurement error. At the same time, in addition to the large range of 165000m, which represents the distance over which the spatial autocorrelation in the main direction become negligible, a minor range of about 55121m is now detected. Related to this, an anisotropy direction of about 39° can be determined, which corresponds to the main axis of spatial continuity. In this context, the semivariogram map now also shows an elliptical shape reflecting a stronger spatial continuity along the major axis.

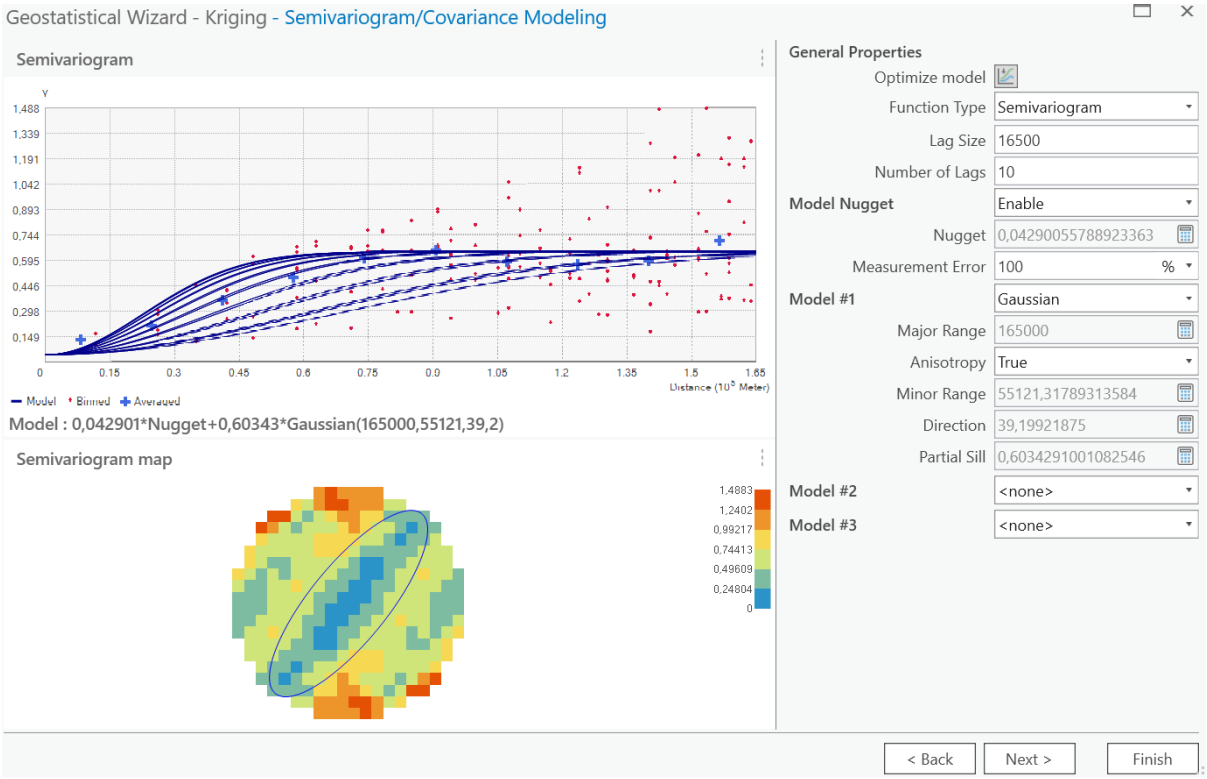

Figure 6. Ansotropic semivariogram based on a Gaussian function for the study area.

9. *Specify an appropriate Searching Neighborhood and justify your choice (parameters: Maximum Neighbors, Minimum Neighbors, Sector Type).*

First, a range ellipse was chosen to ensure that the spatial search reflects the directional structure of the spatial autocorrelation with points that are more strongly correlated in some directions than in others. The prediction is based on up to 20 neighbors, where the minimum number of neighbors is 2, the maximum number of neighbors is 5 and the number of sectors is 4. This number of neighbors provides a balance between prediction stability and local sensitivity. The sector type is defined here as 4 sectors with 45°, which compensates for an uneven point distribution in the study area and thus ensures a more spatially balanced selection of neighbors.

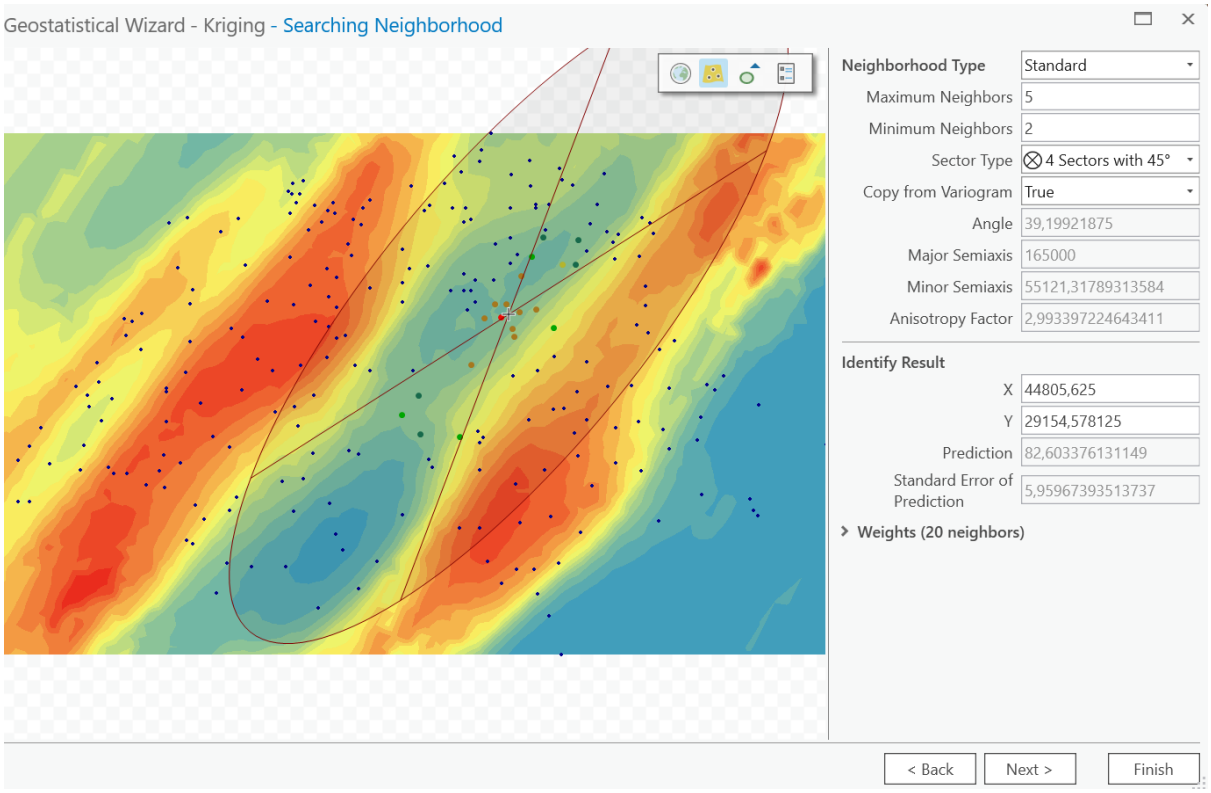

Figure 7. Searching neighborhood with maximum and minimum neighbors as well as sector type.

10. *Carry out a crossvalidation with interpretations (cross validation plots, error parameters). Also provide a brief explanation of cross validation procedures.*

To assess the accuracy and reliability of the kriging interpolation, a cross validation was performed. This involves temporarily removing individual points from the dataset and then using the remaining points to create a new kriging prediction for that particular location. After this, the predicted and actual value can be compared. Once the process has been repeated for all points, several statistics are generated to determine the accuracy.

In the example, the perfect regression line is shown in blue and the actual prediction in gray. While the general trend follows the diagonal, there is an underprediction of higher values, shown here by the mountain peaks. This can also be recognized by a slightly positive mean error, which indicates that the model is overpredicting. At the same time, the distribution of the residuals does not follow a normal distribution, which suggests that some spatial patterns may remain unexplained. This is shown by the relatively high Root-Mean-Square error of around 52.58.

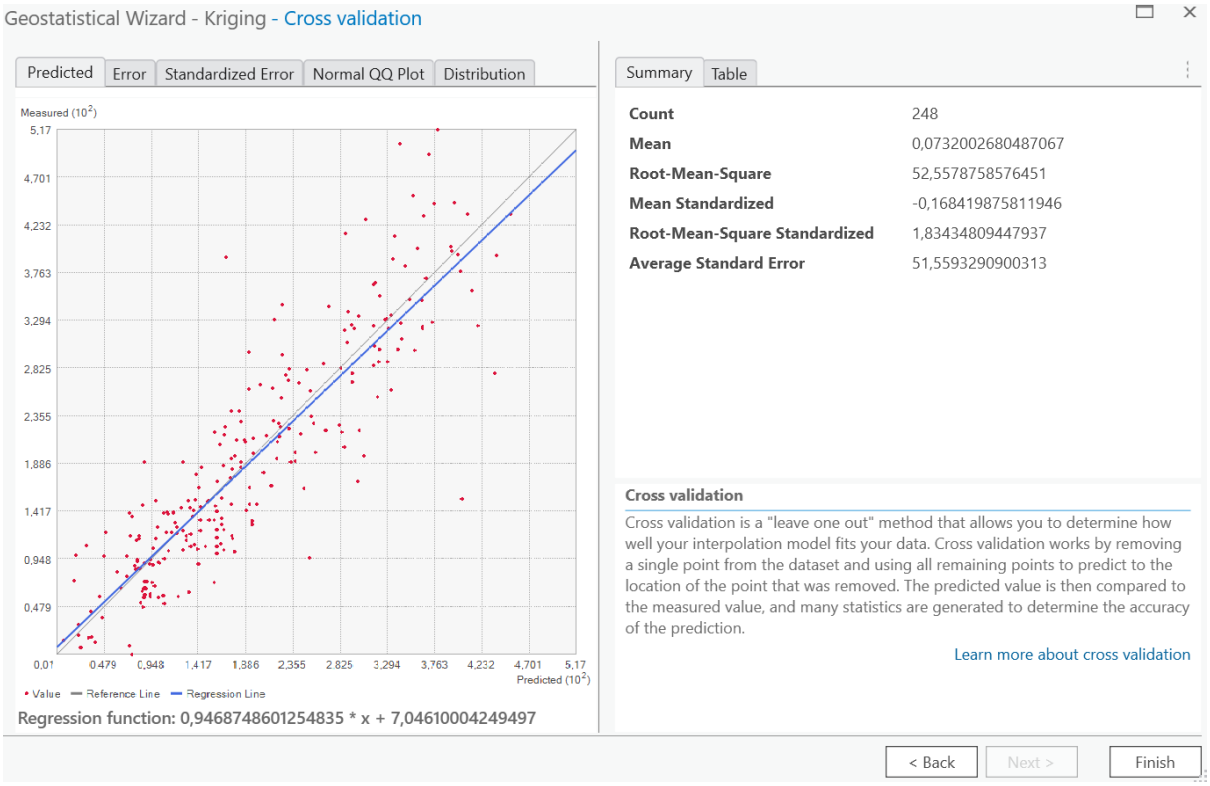

Figure 8. Cross validation results of the kriging interpolation.

11. *Interpret prediction and standard error map: Select a specific prediction value and discuss the uncertainty associated with this prediction based on the standard error. Express uncertainties for this single prediction based on confidence interval and probability.*

In general, the prediction map shows high precipitation values in the mountains and low precipitation values in the valley. The highest precipitation corresponds to the highest peaks in Switzerland. On the other hand, the standard error map shows the lowest standard error, which is a probability density function, in the valley in the north, where there are many data points. In contrast, the east and west in particular have very high standard errors, which are due to a low density of data points.

Comparing the prediction and the corresponding standard errors of a specific value in the valley and on a mountain peak, it is initially noticeable that, compared to the mountain peak, there is not only a lower predicted precipitation, but also a lower standard error of the prediction (see Table 1). However, if these values are related to the confidence interval and the probability, the yearly precipitation in the valley lies between 33 and 41cm with a probability of 68%. Similarly, the yearly precipitation at the selected location at high altitude is between 381 and 457cm with a probability of 68%.

Table 1. Specific prediction values for the valley (left) and the mountain top (right).

Prediction	36,85566
Standard Error of Prediction	4,220391
X Coordinate	-11302,585768
Y Coordinate	-47938,553432

Prediction	418,6595
Standard Error of Prediction	38,098912
X Coordinate	-46891,115526
Y Coordinate	17254,306579

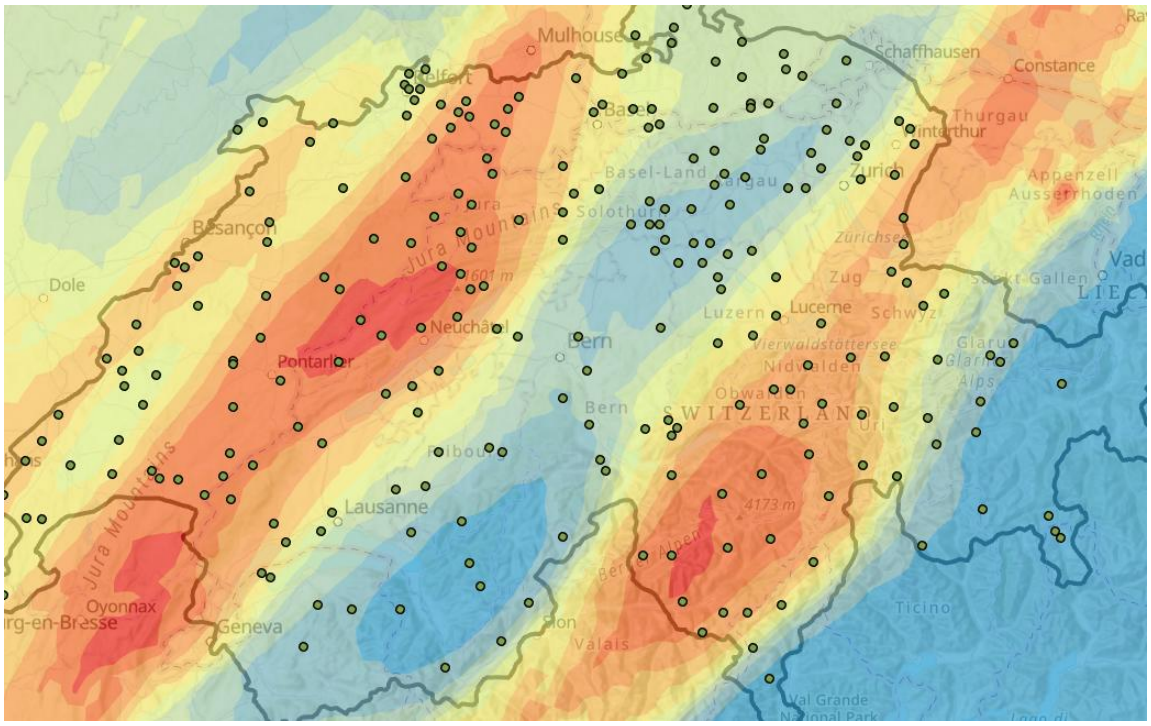

Figure 9. Prediction map after the kriging interpolation.

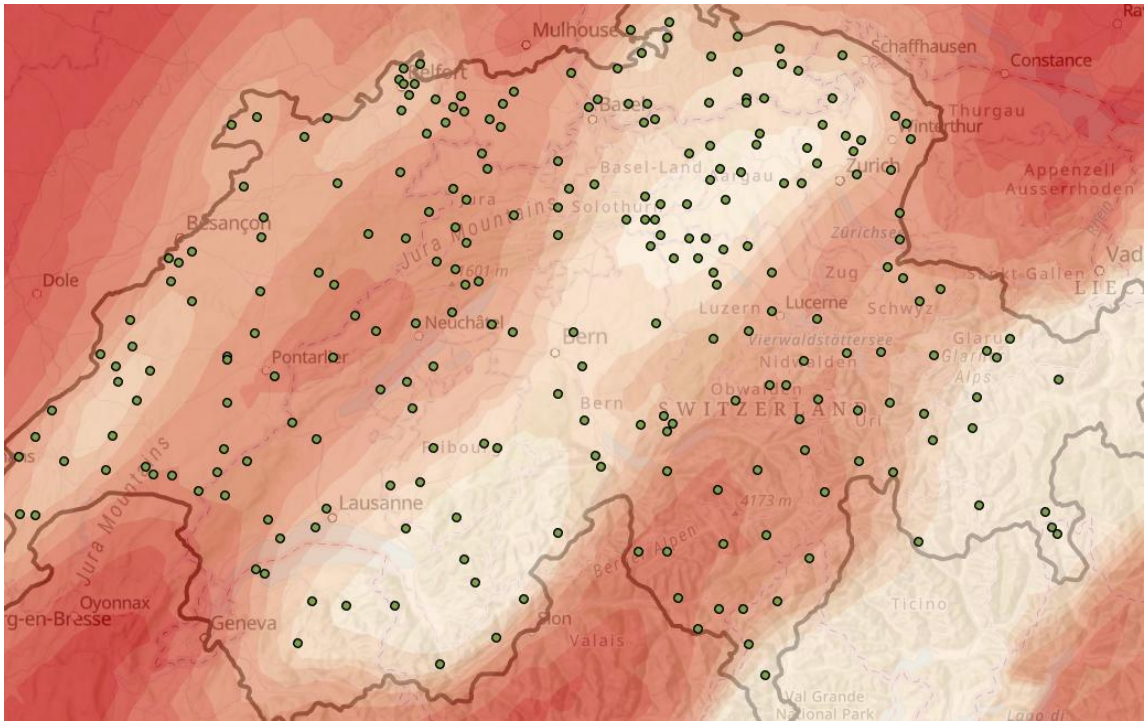

Figure 10. Standard error map after the kriging interpolation.

12. *Briefly discuss the complementary nature of Cross Validation Error und Standard Error. What is model overfitting? Why do overfitted models show small cross validation errors and high average standard errors?*

While the Cross Validation Error refers to the difference between the measured and predicted values, the Standard Error reflects the generic quality of the model. More specifically, the Cross Validation Error shows how the model performs at known locations and is tied to the individual predictions made with the kriging interpolation. On the other hand, the Standard Error shows how well the model performs when repeated, even at unsampled locations. The two metrics therefore complement each other and should both be considered for assessing model quality.

Overfitted models have small Cross Validation Errors and high Standard Errors because the model is too closely tailored to the training data and fails to generalize. Increasing the number of lags results in less data in each lag, which leads to a low Cross Validation Error. At the same time, however, the model cannot reliably interpolate beyond these scarce points, resulting in a high Standard Error. To avoid overfitting, it is therefore recommended to use a reasonable number of lags that ensure a sufficient number of data points per lag.